Playing with truth

Wintein, S.

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Playing with Truth

ISBN: 978-94-6169-193-4
Printed by: Optima Grafische Communicatie, Rotterdam

Playing with Truth

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 10 februari 2012 om 14.15 uur door

Stefan Wintein

geboren op 6 juni 1979 te Middelburg.

Promotiecommissie:

Promotor:	prof. dr. Harrie de Swart
Copromotor:	dr. Reinhard Muskens
Overige leden:	prof. dr. Filip Buekens
	dr. Paul Egré
	prof. dr. Leon Horsten

Preface

This thesis is the final output of a PhD project that started with a research proposal in which I explained how I would study game theoretic concepts using logic. Or something like that. At the end, the thesis involves (some) game theory and (a lot of) logic. But not in a way that is even remotely connected to the research proposal; the topic of this thesis—self-referential truth—was not mentioned in the proposal at all. Sometimes, it is less exciting to do what one said than to say what one did.

I would like to thank my supervisors, Harrie de Swart and Reinhard Muskens, for their continuous support and belief in me and for the useful suggestions they made with respect to my papers.

Filip Buekens, Paul Egré and Leon Horsten, thanks a lot for your willingness to be part of my committee and for your helpful comments with respect to my work.

Kobus and Thijs, thanks for being my "paranimfen"; as first-year students we jointly moved from Middelburg to Tilburg, and so it's nice that you also back me up at the (formal) end of my studies.

I thank my fellow PhD students for the nice chats, discussions, drinks and time we had together in Tilburg.

I am very happy to have many friends, with most of whom I share roots that lead back to Middelburg. Thanks a lot to all of you for the many good times we had, have and will have. Also, thanks for your interest—in whatever state of mind—in my research.

Thanks to my parents and sister for their love and support and the same goes for the in-laws.

Wilma, thanks a lot for being a wonderful and beautiful friend, for being my wife and for being a great mother of Luc and our second son that we're expecting.

Contents

Chapter 1

Introduction

This thesis consists of seven papers that all (but one) revolve around a single topic: that of *self-referential truth*. The papers, though conceptually connected, stand on their own in the sense that they can be read independently of one another. The order in which the papers are put, however, is such that the papers that occur later on in the thesis may refer to results that were established in earlier papers. Although the papers have a formal character, our formal work is motivated by certain philosophical intuitions concerning the notion of truth. Some of the papers hint at these intuitions, often in their introduction or conclusion, but there is no single paper that is devoted to explicating those intuitions as such. The *general outlook* of this introduction discusses, albeit in a leisurely way, the main intuitions that govern our formal work. More generally, the outlook explains how the various papers of this thesis are connected. The general outlook is followed by a section in which we make some comments on possible directions for *future work*. The introduction concludes with an *overview of the thesis*, consisting of the abstracts of the seven papers.

1.1 General outlook

(1) From riddles about truth to theories of truth. In the 1986 movie *Labyrinth*, Sarah has to reach the castle, located in the center of the labyrinth, in order to get back her kidnapped baby brother. On her journey to the castle Sarah faces the following challenge:

> There are two doors and two guards, one who always lies and one who always speaks the truth. One door leads to the castle in the center of the labyrinth and the other door leads to certain death. The guards know which road leads to the castle. The riddle is to find out which door leads to the castle by asking one of the guards a single yes-no question.

The Labyrinth riddle is due to Raymond Smullyan, who invented lots of riddles with the same structure as the Labyrinth riddle. An essential characteristic of Smullyan's riddles is the presence of what he calls *knights*, who always speak the truth, and *knaves*, who always lie. Another such characteristic is that, in order to solve the riddle, you have to address *yes-no questions* to creatures that may

be knights or knaves; the difficulty is that, typically, you do not know whether you are facing a knight or a knave. Riddles which have those characteristics, we call *riddles about truth*.

Facing the two guards, Sarah manages to continue her journey successfully by addressing the following yes-no question to the guard of the left door:

> Would *he*—Sarah points at the guard of the right door—tell me that *this door*—Sarah points at the left door—leads to the castle?

Upon hearing Sarah's question, the left guard needs some time to deliberate, but then he answers the question with 'yes'. From this answer, Sarah concludes that the right door leads to the castle and, before she continues her journey to the castle via the right road, she spends some time on explaining the perplexed guards the rationale of her solution.

In a broad sense, the rationale of solutions to riddles about truth, of which the Labyrinth riddle is an example, is the topic of Section 2 and 3. For sure, Sarah's reasoning is correct, but, so one may ask, in virtue of which principles? In order to answer that question, Section 2 develops a *framework* in which riddles about truth can be formalized. The relevance of the development of such a framework is partly explained by the light that it sheds on another riddle about truth, called the *Hardest Logic Puzzle Ever* ($HLPE$) by George Boolos, that attracted quite some attention in the academic literature. $HLPE$ is formulated as follows:

> The Puzzle: Three gods A, B and C are called, in some order, True, False, and Random. True always speaks truly, False always speaks falsely, but whether Random speaks truly or falsely is a completely *random* matter. Your task is to determine the identities of A, B, and C by asking three yes-no questions; each question must be put to exactly one god. The gods understand English, but will answer all questions in their own language, in which the words for 'yes' and 'no' are 'da' and 'ja' in some order. *You do not know which word means which.* (Boolos [9, p62])

The solutions to (versions of) $HLPE$ as given by Boolos and others (cf. [47], [43], [53], [56]) share a common characteristic; they are all stated in natural language. Our framework provides a formal basis for comparison of natural language solutions to riddles about truth in general, and to $HLPE$ in particular. Besides representing natural language solutions, our framework can be used to make precise certain claims that the authors make *about* their solutions. For instance, Tim Roberts [47] criticizes Boolos' solution to $HLPE$ as being "too complicated", after which he comes up with an alterative solution to $HLPE$ that, so he claims, is less complicated. Upon representing their respective solutions in our framework, however, we see that there is an important sense in which it is Roberts' solution that is more complicated.

In formalizing riddles about truth, the crucial question is how we should model the behavior of knights (and knaves). As a knight always speaks *truly*, the behavior of a knight is obviously connected to the notion of *truth*. It is no surprise then, that our formalizations of the behavior of a knight rely on formal *theories of truth*. We feel that, besides shedding light on solutions to $HLPE$, another interesting feature of our formalizations of riddles about truth is the

connection they bear with theories of truth.

In Section 4, 5, 6 and 7, we are not concerned with riddles about truth[1], but rather with (formal) theories of truth. There are close connections between the theories of truth that are considered in this thesis. On the technical side, the connections can be understood from the relations between the *frameworks* that we develop to define and study theories of truth: in Section 4, we develop *assertoric semantics*, in Section 5, we develop the *method of closure games* and in Section 7, we develop the *strict-tolerant calculus*. The development of our three frameworks is guided by a single rationale, which may be called a *deflationary, assertoric conception of truth*. Before we turn to this rationale, we introduce the notions of a *quotational truth language*, a *ground model* and of a *theory of truth*, which play an important role in all (but one) of the sections of this thesis. After our presentation of the main ingredients of a deflationary, assertoric conception of truth, we explain how it is cashed out in terms of the three—interrelated—formal frameworks that constitute the heart of this thesis.

(2) Quotational truth languages and ground models. A *quotational language* is a first order language L with some additional structure, guaranteeing that L has the means to refer to all of its sentences via *quotational names*. A quotational language generalizes and formalizes a natural language platitude pertaining to quoting: 'snow is white' refers to the following sentence: snow is white. Similarly, each sentence σ of a quotational language L has a quotational name, $[\sigma]$, in L. A *quotational truth language* L_T is a quotational language with a distinguished predicate symbol, T, whose intended interpretation is that of a *truth predicate*. That is, $T(x)$ is to be read as 'x is true'. A *ground model* for (a quotational truth language) L_T is a classical interpretation of L, the truth-free fragment of L_T. Intuitively, a ground model represents "the non-semantic states of affairs". A ground model M equips L with a $\mathcal{C}$*lassical semantic valuation*, which we will denote as $\mathcal{C}_M : Sen(L) \to \{1, 0\}$. By a *theory of truth*, we will mean...

> ...a theory that purports to explain for a first-order language L_T what sentences are assertible in a [ground] model M. (Gupta, [23, p19])

(3) Explanatory versus expressive function of truth. Suppose that you confront Sarah with the following formulation of the Labyrinth riddle:

> There are two doors and two guards. One guard is a *knight*, the other one a *knave*. One door leads to the castle in the center of the labyrinth and the other door leads to certain death. The guards know which road leads to the castle. The riddle is to find out which door leads to the castle by asking one of the guards a single yes-no question.

Suppose that Sarah is not familiar with Smullyan's work; Sarah has never heard of knights (and knaves). Hence, she asks for an explanation[2]: what makes a knight? As a response to Sarah's question, we offer the following *explanation*:

[1] Although Section 4 is concerned with *query problems* which, in a sense, are abstractions of riddles about truth.

[2] Concerning Sarah's question, see also *What makes a knight?* [59], a paper that didn't make it to this thesis.

> A *knight* is a creature which behaves as follows: a knight answers a yes-no question with 'yes' just in case the associated declarative sentence is *true*, whereas he answers with 'no' otherwise.

As an example, the declarative sentence associated with the yes-no question 'is snow white?' is 'snow is white'. As 'snow is white' is true, a knight answers the associated yes-no question with 'yes'. Further, we explain Sarah how a knave behaves, after which she successfully continues her journey.

In terms of L_T, our explanation of a knight's behavior can be phrased as follows:

$$\forall x\ (\mathcal{A}(x, c_y) \leftrightarrow T(x)) \wedge (\mathcal{A}(x, c_n) \leftrightarrow \neg T(x)) \tag{1.1}$$

Here $\mathcal{A}(x, y)$ reads as "the $\mathcal{A}$nswer of the knight to x is y", c_y and c_n stand for 'yes' and 'no' respectively, $\leftrightarrow$ is the material biconditional and $\wedge$ expresses conjunction.

In explaining, to Sarah, what a knight is, we use the notion of truth. Hence, our toy example suggests that truth can play an *explanatory* role or function. However, Paul Horwich [28] famously argued that truth *never* plays an explanatory role. Rather, the sole function of our truth predicate is an *expressive one*. Although we are sympathetic to this particular claim of Horwich, we have certain qualms with his conception of truth, *minimalism*, to which it is wedded. We first sketch a typical "Horwichean" analysis of our explanation to Sarah, after which we explain on which points we agree and disagree with Horwich.

(4) A minimalist explanation of a knight's behavior. According to Paul Horwich, the expressive function of the truth predicate is, in an important sense, *all there is to the notion of truth*. It is generally acknowledged that an important function of our truth predicate is an expressive one, allowing us to make assertions for which the truth predicate is indispensable. Examples of such assertions are *generalizations* (1.2) and *blind ascriptions* (1.3) .

Everything the Pope says *ex cathedra* is true. (1.2)

Einstein's first theorem is true. (1.3)

That truth plays an expressive function is uncontroversial. That the function of truth is *exhausted* by its expressive function, is a thought that is associated with, amongst others, Horwich conception of truth, which is called *minimalism*. According to Horwich, the sole function of truth is an expressive one, and truth can play its expressive role due to our 'unconditional acceptance of the (unproblematic) instances of the equivalence schema'. Or, to quote Horwich:

> The entire conceptual and theoretical role of truth may be explained on the basis of all uncontroversial instances of the equivalence schema: it is true that p if and only if p.
>
> (Horwich, [28, p5])

Horwich states the *equivalence schema*, consisting of the *Tarski-biconditionals*, in terms of propositions, we do so in terms of sentences[3]. By exploiting quote names, we may state the equivalence schema **ES** as follows:

$$\mathbf{ES}: T([\sigma]) \leftrightarrow \sigma$$

[3] More generally, the main results of this thesis are obtained by *modeling* truth as a predicate T applying to the sentences of a formal language. Now, it may very well be that, for subtle

Thus, according to Horwich's minimalism, the sole function of truth is an expressive one and truth has this function in virtue of our commitment to all instances of **ES**. Let's now turn to a minimalist analysis of the role of truth in the explanation of a knight's behavior.

A knight is a creature that answers 'is snow white?' with 'yes' because snow is white. Further, a knight answers 'Is the number of particles in the universe odd?' with 'yes' just in case the number of particles in the universe is odd. More generally, a knight answers 's?' with 'yes' just in case s. More formally, the behavior of a knight is specified by the following axiom schema:

$$(\mathcal{A}([\sigma], c_y) \leftrightarrow \sigma) \wedge (\mathcal{A}([\sigma], c_n) \leftrightarrow \neg\sigma) \tag{1.4}$$

In contrast to (1.1), the account of a knight's behavior as in (1.4) does not invoke the notion of truth. As (1.4) is a "truth-free explanation" of a knight's behavior, we see that the role of truth in such an account is not, fundamentally, an explanatory one. Being finite creatures though, we cannot communicate how a knight behaves by asserting all of the infinitely many instances of (1.4). However, by exploiting the truth predicate in combination with a universal quantifier, we can, albeit indirectly, assert all instances of (1.4) by asserting (1.1). Thus, the truth predicate plays an *expressive function* in the explanation of a knight's behavior. Moreover, to realize this expressive function, **ES** plays a crucial role:

1. $\forall x\ (\mathcal{A}(x, c_y) \leftrightarrow T(x)) \wedge (\mathcal{A}(x, c_n) \leftrightarrow \neg T(x))$
2. $(\mathcal{A}([\sigma], c_y) \leftrightarrow T([\sigma])) \wedge (\mathcal{A}([\sigma], c_n) \leftrightarrow \neg T([\sigma]))$ (from 1)
3. $\mathcal{A}([\sigma], c_y) \leftrightarrow T([\sigma])$ (from 2)
4. $T([\sigma]) \leftrightarrow \sigma$ (from **ES**)
5. $\mathcal{A}([\sigma], c_y) \leftrightarrow \sigma$ (from 3, 4)
6. $\mathcal{A}([\sigma], c_n) \leftrightarrow \neg\sigma$ (by taking steps similar to 2,3,4)
7. $(\mathcal{A}([\sigma], c_y) \leftrightarrow \sigma) \wedge (\mathcal{A}([\sigma], c_n) \leftrightarrow \neg\sigma)$ (from 5, 6)

So, (1.1) allows us, in combination with **ES** (and classical logic), to explain the behavior of a knight by (indirectly) asserting all instances of (1.4); by asserting (1.1) we commit ourselves to all instances of (1.4). However, in "the explanation itself", i.e., in (1.4), the notion of truth does not play a role.

To be sure, this was just a toy example. But it has the same structure as the examples that are used by Horwich [28] to argue for his claim that the sole function of truth is an expressive one. In the third chapter of [28], called *The Explanatory Role of the Concept of Truth*, Horwich seeks to rebut the following claim that he ascribes to, amongst others, Putnam [42] and Field [14].

philosophical reasons, the truth predicate of *our language*, '... is true', must be understood as applying to propositions (or beliefs) rather than to sentences. Then again, it seems that these propositions are expressible via sentences and so our truth predicate applies at least derivatively to sentences. Further, given our formal set-up, it is far more convenient to talk about the truth predicate (be it ours or a formal one) as applying to sentences. Hence, we will do so.

> Truth has certain characteristic effects and causes. For example, true beliefs tend to facilitate the achievement of practical goals. General laws such as this call for explanation in terms of the nature of truth. (Horwich, [28, p44])

Horwich rebuts the claim by showing that:

> As we shall see [in various examples with the same structure as our toy example], truth does indeed enter into explanatory principles, but their validity may be understood from within the minimal theory [i.e., the instances of the equivalence schema].
>
> (Horwich, [28, p44])

That is, by analyzing cases where truth is purported to play an explanatory role, we see that, in fact, the function of truth is to allow us to assert, albeit indirectly via **ES** and logic, certain "truth-free sentences".

(5) Agreeing with Horwich I am sympathetic to at least two ideas of Horwich's conception of truth.

First, I take it that the sole function of our *truth predicate* is indeed an expressive one. I will not argue for this claim though, but rather, build upon it. When talking about *truth* though, some care is needed. For, so one may ask, is it the *property* of 'being true' that is under consideration, the *concept* of truth, or the *predicate* ' . . . is true'? According to some authors, Horwich's *minimalism* is correct in its analysis of the truth predicate, but it fails to account for the role of, in particular, the *concept of truth*. For instance, Filip Buekens [10] takes it that Horwich's minimalism says the right things about the truth predicate, but argues that it doesn't account for the explanatory role of the truth concept in rational reconstructions of what we mean and say. In this thesis, talk about truth is to be understood as talk about the truth predicate. We take it that the sole function of the truth predicate is an expressive one. By doing so, we do not exclude the possibility that, say, the concept of truth fulfills a function in our cognitive lives that cannot be understood in terms of the expressive function of the truth predicate. However, this thesis is not concerned with such possibilities.

Second I am sympathetic to Horwich's idea that, in an important sense, "truth has no underlying nature". To illustrate the latter claim, consider the following sentences:

$$\text{Snow is white.} \tag{1.5}$$

$$\text{Theft is (morally) wrong.} \tag{1.6}$$

Although their subject matters differ widely, both sentences are true. According to *substantial* theories of truth, (1.5) and (1.6) must share a property, P, in virtue of which they are true. Substantial theories of truth have explicated P as, e.g.,"correspondence to the facts" (correspondence theory), or "membership of a coherent set of beliefs" (coherence theory). When "truth has no underlying nature" the substantial theories fail, because truth is a *primitive notion*, i.e., there is no P which explains what all truths have in common. In this thesis, we do not argue for the claim that truth is a primitive notion. Rather, we build upon the assumption that truth is best understood as such.

A conception of truth which understands truth as a primitive notion (which only plays an expressive function) is called a *deflationary conception of truth*. In this thesis, we will develop a deflationary conception of truth which differs from Horwich's minimalism in the following two, related, ways. First, truth plays its expressive function not in virtue of **ES**, but rather, in virtue of its *transparency*, as explained in **(10)** below. Second, the primitivity of truth is not to be understood in terms of our 'unconditional acceptance of the (unproblematic) instances of the equivalence schema', but rather, in terms of our 'unconditional acceptance of the assertoric rules of truth', as explained immediately below.

(6) Disagreeing with Horwich: assertoric rules of L_T. Although I am sympathetic to the *deflationary spirit* of Horwich's minimalism, the equivalence schema will not have a major role to play in this thesis. In this thesis, the leading role is not for the equivalence schema, but rather for the *assertoric rules of truth* which, in this thesis[4], will be formulated according to the following schema:

$$\frac{A_{T([\sigma])}}{A_\sigma} \qquad \frac{D_{T([\sigma])}}{D_\sigma}$$

The signs A and D are associated with an assertion and denial respectively. Before we comment on the association, we remark that the assertoric rules for truth allow us to understand truth as being on a par with the other logical connectives, in the sense that these can also be understood via their assertoric rules. For instance, here are the rules for negation and conjunction:

$$\frac{A_{\neg\alpha}}{D_\alpha} \quad \frac{D_{\neg\alpha}}{A_\alpha} \qquad \frac{A_{\alpha\wedge\beta}}{A_\alpha, A_\beta} \quad \frac{D_{\alpha\wedge\beta}}{D_\alpha \mid D_\beta}$$

In this thesis, we distinguish two distinct readings of the assertoric rules of L_T, which we call the *commitment* and *entitlement* reading respectively. The two readings of the assertoric rules differ in the meaning that they attach to the signs A and D.

Commitment reading On the commitment reading, A_σ and D_σ indicate, respectively, a commitment to an assertion of σ and a commitment to a denial of σ. We take it that there are two ways in which one can become committed to an assertion (denial) of σ. In a *direct manner*, via an outright assertion (denial) of σ, or in an *indirect manner*, i.e., as a function of (previous) outright assertoric actions and the assertoric rules. For instance, I become indirectly committed to an assertion of α via an outright assertion of $\alpha \wedge \beta$ and the assertion rule of $\wedge$, or conversely, I become indirectly committed to an assertion of $\alpha \wedge \beta$ via (previous) outright assertions of α and β and the assertion rule of $\wedge$.

[4]In Section 7, superscripts will be added to indicate the sense (*strict* or *tolerant*) in which a sentence is asserted or denied.

We take it that (but see Section 6) the commitment reading of the assertoric rules is valid *in both directions*[5]. For instance: one is committed to an assertion (denial) of $T([\sigma])$ iff one is committed to assertion (denial) of σ. Similar for the other assertoric rules. A remark is in order here. Consider the following sentence:

JC: Ceasar had exaclty 12 hairs on his big toe when he crossed the Rubicon.

Suppose that I assert $JC \vee \neg JC$, which seems perfectly fine, as either Ceasar had or didn't had 12 hairs on his big toe when he crossed the Rubicon. But then, under the commitment reading of the assertoric rules, by asserting $JC \vee \neg JC$ I become (indirectly) committed to an assertion of JC or to an assertion of $\neg JC$. But this seems absurd, as I do not (and no one does) *know* whether JC or $\neg JC$. Hence, so one may ask, which sense of assertion and denial do we model by the *assertoric* interpretation of the AD rules. A quick and dirty answer is that we are modeling *idealized* assertbility, that is, assertibility for an agent who has full knowledge of all non semantic facts as represented by the ground model. This approach is explicitly taken in Section 4, which considers the assertoric actions of an *oracle*, which is omniscient in the sense alluded to. In Section 6 though, we observe that the reaction to JC depends on the knowledge account of assertion, and that this account is not indisputable. There, we suggest that it is possible to make sense of our reading of the assertoric rules along the lines of the truth account of assertion as developed by Weiner [55]. Further, Section 6 illustrates (via a rudimentary example) that the *techniques* that are developed in this thesis to model assertoric norms can be adapted to account for the knowledge account of assertion as well.

Entitlement reading On the entitlement reading, A_σ and D_σ indicate, respectively, an entitlement to an assertion of σ and an entitlement to a denial of σ. Whether or not one is entitled to assert (deny) σ depends, in general, on the *assertoric norm* under consideration.

The notion of an assertoric norm will play a crucial role in this thesis, as will be explained immediately below. Let us note that, in contrast to the commitment reading, we do *not* take the entitlement reading of the assertoric rules to be (generally) valid in both directions—even in situations of "full knowledge of non semantic facts"—as will be explained in **(14)** below.

(7) Assertoric norms It is generally acknowledged that our assertoric practice, i.e., our practice of asserting and denying sentences, is a *rule based* and (hence) *normative* practice. What is not generally acknowledged is how the norm that governs this practice should be understood. Typically, a discussion of an assertoric norm consists of a specification of circumstances under which it is *allowed to assert* a sentence[6]. Here are some proposed assertoric norms:

- *Knowledge norm:* one is allowed to assert σ only if one knows σ.

[5]The directions alluded to in the notions *valid in upwards direction*, *valid in downwards direction* and *valid in both directions* are explained in accordance with the graphical representation of the assertoric rules.

[6]Typically, it is taken for granted that to deny a sentence is to assert its negation. On this view, it suffices to specify norms for assertion only. We'll return to this view in more detail below.

- *Truth norm:* one is allowed to assert σ only if σ is true.
- *Warrant norm:* one is allowed to assert σ only if one has "sufficient" warrant for σ.

There is an important distinction between the Knowledge, Truth and Warrant norm on the one hand, and the assertoric norms that will be considered in this thesis, on the other. We may say that the Knowledge, Truth and Warrant norm are *Prussian norms*. Just like the Prussian conception of law, the norms focus upon what one is allowed to do, by listing those actions that are to be deemed "legal" and by specifying that *everything is forbidden which is not explicitly permitted*. As we will see, our assertoric norms focus upon what one is forbidden to do, by listing those actions that are to be deemed "a crime" and by then specifying that *everything is permitted that is not explicitly forbidden*[7]. In this sense, our norms resemble the English conception of law, and we may say that the norms that will be considered in this thesis are *English norms*. Another distinction between the Knowledge, Truth and Warrant norm and the assertoric norms of this thesis, is that the latter have a more formal character. To illustrate both distinctions at once, let M be a ground model—remember that, intuitively, M specifies the "non-semantic states of affairs", and that it gives rise to a classical valuation $\mathcal{C}_M$ of L—and consider the following assertoric norm which will have a major role to play in this thesis[8].

The strict norm. It is (explicitly) *forbidden* to assert (deny) σ just is case, by asserting (denying) σ, you become *committed*, via the assertoric rules, to:
1) an assertion of an atomic truth-free sentence α of L with $\mathcal{C}_M(\alpha) = 0$, or
2) a denial of an atomic truth-free sentence α of L with $\mathcal{C}_M(\alpha) = 1$, or
3) an assertion and a denial of an arbitrary sentence of L_T.
Whenever it is not (explicitly) forbidden to assert (deny) σ, it is *allowed* to assert (deny) σ.

In a catchy slogan, the strict norm may be stated as follows: *thou shalt respect the world and thou shalt not contradict thyself*. The interpretation of the third condition is clear; as assertion and denial are mutually exclusive, one cannot live up to commitments which specify that you are committed to an assertion and a denial of the same sentence[9]. The interpretation of the first two conditions relies on the interpretation of $\mathcal{C}_M$. In common parlance, $\mathcal{C}_M(\alpha) = 1$ and $\mathcal{C}_M(\alpha) = 0$ abbreviate 'α is true' and 'α is false' respectively. The common parlance may also be applied in the present setting, but please note that, in doing so, the truth predicate plays an expressive, and not an explanatory function. Thus, we do not explain assertion and denial in terms of truth. More concretely, an example of a truth free-sentence is 'snow is white'. It is beyond doubt that we are allowed to assert 'snow is white' and it is such information that is captured, via a ground model, by the classical valuation $\mathcal{C}_M$. For the purposes of this thesis, such a (minimal) interpretation of $\mathcal{C}_M$ suffices. According to this interpretation, we do not have to (and do not want to) say that 'snow is white' is assertible

[7] The Prussian / English terminology is taken from van Fraassen [54], who uses it to draw a distinction between two accounts of rationality.

[8] See in particular Section 4, 5 and 7

[9] See Section 7 though, were we consider the thought that, e.g., *Liar sentences*, are *both* tolerantly assertible and deniable.

because it is true. Likewise, we do not have to (and do not want to) say that 'snow is white' is true *because* it is assertible. Saying that 'snow is white' is assertible *because* it is true is tantamount to ascribing an explanatory role to truth. Saying that 'snow is white' true *because* it is assertible is tantamount to ascribing an "underlying nature" to truth. As we explained in **(5)**, we agree with Horwich that such ascriptions should be avoided.

(8) An assertoric conception of truth. We understand the notion of truth, on a par with the (other) logical connectives, in terms of its *assertoric rules*. Therefore, we will say that we advocate an *assertoric conception* of truth. As such, an assertoric conception of truth is not committed to the strict norm, although the norm will play a major role in this thesis. Further details of our assertoric conception of truth will be filled in below. We will start though, by explaining how the phenomenon of self-referential truth testifies that accepting the assertoric rules for truth significantly differs from accepting the instances of **ES**.

(9) Self-reference: how the assertoric truth rules and ES come apart. There is a fundamental distinction between accepting the *assertoric rules* of truth, on the one hand, and the equivalence schema **ES**, on the other. The fundamental character of the distinction may not be clear at first sight. For, can't we say that, in asserting the Tarski-biconditional ' 'snow is white' is true if and only if snow is white', we thereby express that we *accept* the (assertoric) truth rules pertaining to the sentence 'snow is white'? If such a relation holds in general, then, to call the distinction between **ES** and the assertoric rules *fundamental* is to make too much fuss.

Very well, but the relation between the Tarski-biconditional and the assertoric truth rule of a sentence is not always as in the case of 'snow is white'. The phenomenon of *self-referential truth*, which is the central topic of this thesis, testifies that we may accept the assertoric truth rules for any sentence, while there are sentences for which we are reluctant to express this acceptance via a Tarski-biconditional. To illustrate that the rules for truth and the equivalence schema may come apart in the sense alluded to, consider :

$$\text{sentence (1.7) is not true.} \tag{1.7}$$

We may say that (1.7) says, of itself, that is not true. Sentences which assert their own untruth, such as (1.7), are called *Liar sentences*. In terms of L_T, a Liar sentence will be represented as the sentence $\neg T(\lambda)$, where λ is specified to denote[10] $\neg T(\lambda)$. We will now use $\neg T(\lambda)$ to explain, in terms of our framework, the distinction between accepting the assertoric truth rules and accepting **ES**.

Suppose that you assert $\neg T(\lambda)$. By doing so, you take up assertoric commitments in accordance with the assertoric rules; in particular, you take up the commitment to deny $T(\lambda)$—in accordance with assertoric rules for negation. Further, the commitment to deny $T(\lambda)$ brings with it further assertoric commitments, which, on their turn, ... By asserting $\neg T(\lambda)$, we take up assertoric commitments that, according to the strict norm, *we cannot live up to*, for:

[10] In fact, this thesis exploits two distinct ways to express that 'λ denotes $\neg T(\lambda)$'. In Section 7, we do so in our object language by exploiting an identity sign $\approx$ via the sentence: $\lambda \approx [\neg T(\lambda)]$. In Section 2,3,4,5 and 6 we do so via a denotation function I which is such that $I(\lambda) = \neg T(\lambda)$.

1. $A_{\neg T(\lambda)}$ (asserting $\neg T(\lambda)$)
2. $D_{T(\lambda)}$ (from 1 and the rule $A_{\neg}$)
3. $D_{T([\neg T(\lambda)])}$ (from 2 by substitution, as λ denotes $\neg T(\lambda)$)
4. $D_{\neg T(\lambda)}$ (from 3 and the rule D_T)

Hence, by asserting $\neg T(\lambda)$ you become, via the assertoric rules, committed to a denial of $\neg T(\lambda)$. Accordingly, an assertion of $\neg T(\lambda)$ is forbidden. Dually, a denial of $\neg T(\lambda)$ is also seen to be forbidden. Hence, one should neither assert nor deny $\neg T(\lambda)$. To arrive at the judgement that the Liar is neither assertible nor deniable, we applied the assertoric truth rules pertaining to the Liar. Although we accept these rules, we cannot express that we do so via the associated Tarski-biconditional of (1.8).

$$T([\neg T(\lambda)]) \leftrightarrow \neg T(\lambda) \tag{1.8}$$

By asserting or denying (1.8), we likewise take up commitments that we cannot live up to according to the strict norm, as the reader may verify for himself. So, (1.8) shares its assertoric status with the Liar and so, in particular, it is not allowed—under the strict norm—to assert (1.8). Hence, our acceptance of the assertoric truth rules pertaining to the Liar cannot be expressed via an assertion of the associated instance of the equivalence schema. Accepting the assertoric rules for truth thus differs from accepting all instances of **ES**.

To be sure, Horwich reserves a central role for all *uncontroversial* instances of the equivalence schema, and (1.8) is the canonical example of a *controversial* instance of **ES**: including (1.8) in **ES** produces an inconsistent theory from which—given Horwich's commitment to classical logic—everything follows. So, Horwich's minimalism needs to be backed up by a sieve that can demarcate the controversial from the uncontroversial instances of **ES**. There may very well be such a sieve[11]. Then again, we feel that the introduction of such a sieve is an *ad hoc* move. However, the deflationary spirit (truth is a primitive notion whose sole function is an expressive one) of Horwich's position does not seem to rely on his insistence on **ES**. More concretely, we can understand the primitivity of truth in terms of its assertoric rules and we can explain its expressive function in terms of its *transparency*. To illustrate this last remark, we return to the explanation of a knight's behavior.

(10) Transparency and the expressive function of truth. Recall that we take a *theory of truth* to be...

> ...a theory that purports to explain for a first-order language L_T what sentences are assertible in a [ground] model M. (Gupta, [23, p19])

In a considerable part of this thesis (Section 4, 5, 7), we will develop methods to define (construct) theories of truth. As we will see later on, the methods wear the assertoric conception of truth on their sleeves: roughly, a theory of truth is defined by specifying a formal English assertoric norm which, in combination with the assertoric rules, tells us which sentences of L_T are assertible (and or

[11] Horwich sometimes suggests that all and only *grounded*—a notion that is discussed at various places in this thesis—instances of **ES** are uncontroversial.

deniable). All the theories of truth that will be defined in this thesis share the characteristic that, according to these theories, truth is a *transparent* notion. Many authors (most prominently Field [15] and Beall [5]) have argued that truth can play its expressive function in virtue of this transparency. The following quote explains the notion of transparency and, also, illustrates its central importance in Beall's conception of truth.

> This book has a single aim: to concisely lay out and defend a simple, modest approach to *transparent truth* and its inevitable paradoxes, where transparent truth is entirely 'see-through' truth, a notion of truth such that x *is true* and x are intersubstitutable in all (non-opaque) contexts, for all (meaningful, declarative) sentences x of our language. (Beall, [5, p.vii])

In our terminology, truth's transparency is described as follows. Let ϕ be a sentence of L_T which contains $T([\sigma])$ as a sub sentence and let ϕ' result from ϕ by replacing (one or more occurrences of) $T([\sigma])$ with σ. A theory of truth dictates that truth is transparent if, for all ϕ and ϕ' that are related as indicated, the assertoric status of ϕ is the same as that of ϕ'. Before we discuss the relation between transparent truth and the assertoric conception of truth in more detail, we first illustrate the sense in which the transparency of truth ensures that truth can play its expressive function.

Remember that, for the sake of argument, we initially took our explanation of a knight's behavior to suggest that truth can play an explanatory function. Then, we illustrated how Horwich would invoke his *minimalist* conception of truth to rebut this suggestion. On behalf of Horwich, we argued that the function of truth in the explanation of a knight's behavior is to let us take up a commitment to all instances of (1.4) via an assertion of (1.1). Moreover, we saw how this function could be realized (or explained) via (classical logic and) the acceptance of the equivalence schema **ES**. However, the function (of taking us to (1.4) via (1.1)) can also be realized by means of a transparent notion of truth. Here is how:

1. $\forall x\ (\mathcal{A}(x, c_y) \leftrightarrow T(x)) \wedge (\mathcal{A}(x, c_n) \leftrightarrow \neg T(x))$
2. $(\mathcal{A}([\sigma], c_y) \leftrightarrow T([\sigma])) \wedge (\mathcal{A}([\sigma], c_n) \leftrightarrow \neg T([\sigma]))$ (from 1)
3. $(\mathcal{A}([\sigma], c_y) \leftrightarrow \sigma) \wedge (\mathcal{A}([\sigma], c_n) \leftrightarrow \neg\sigma)$ (from 2 and truth's transparency)

More generally, the notion of transparent truth allows us[12] to account for truth's expressive function without relying on the uncontroversial instances of **ES**.

(11) Transparent truth and Strong Kleene fixed point valuations. The transparency of truth is a property of so called *Strong Kleene (SK) fixed point valuations*. With M a ground model, $V_M : Sen(L_T) \to \{0, \frac{1}{2}, 1\}$ is a SK fixed point valuation over M just in case:

[12]Note that the transparency of truth does not guarantee that all (intuitively plausible) truth involving generalizations are assertible according to a theory of truth. For instance, according to Kripke's minimal fixed point theory of truth (which satisfies the transparency of truth), a statement like 'for any sentences α and β, their disjunction is true iff α is true or β is true' is not assertible.

- $V_M(\sigma) = \mathcal{C}_M(\sigma)$ for all $\sigma \in Sen(L)$
  V_M *respects the ground model* M*.*

- $V_M(T(\overline{\sigma})) = V_M(\sigma)$, whenever $\overline{\sigma}$ denotes σ in M.
  V_M *respects the identity of truth*

- $V_M(\neg\sigma) = 1 - V_M(\sigma)$
  $V_M(\alpha \wedge \beta) = \min\{V_M(\alpha), V_M(\beta)\}$, $V_M(\alpha \vee \beta) = \max\{V_M(\alpha), V_M(\beta)\}$
  $\exists$ and $\forall$ behave as generalized $\vee$ and $\wedge$ respectively.

  V_M *dictates that the logical connectives have a Strong Kleene semantics.*

Thus formulated, a SK fixed point valuation V_M is not a theory of truth in the sense of this thesis, as the values $0, \frac{1}{2}$ and 1 as such do not inform us about the assertoric status of the L_T sentences. Sentences with value 1 will be assertible (and not deniable), no doubt. Sentences with value 0 will be deniable (and not assertible), for sure. The interpretation of $\frac{1}{2}$ though, has been much disputed. Before we comment on that dispute, we make a couple of remarks on the abstract notion of a SK fixed point valuation, which will be useful later on.

Multiple SK fixed point valuations over M*.* There are various SK fixed point valuations over a fixed ground model M. The Liar sentence though, has to be valuated as $\frac{1}{2}$ in any SK fixed point valuation; allotting it a value of 1 or 0 is incompatible with the Strong Kleene behavior of negation and the identity of truth. To illustrate the existence of multiple SK fixed point valuations over M, we turn to the "benign cousin" of the Liar. Consider:

$$\text{sentence (1.9) is true.} \tag{1.9}$$

Sentence (1.9) says, of itself, that is true. Sentences which assert their own truth, such as (1.9), are called *Truthtellers*. In terms of L_T, a Truthteller will be represented as the sentence $T(\tau)$, where τ is specified to denote $T(\tau)$. Clearly, the valuation of the Truthteller does not depend on the ground model M. Then, given the structure of the Truthteller, it is not hard to see that there are three distinct types of SK fixed point valuations over M: those that valuate $T(\tau)$ as $0, \frac{1}{2}$ and 1 respectively. We will use $\mathbf{FP}(L_T, M)$ to denote the set of all SK fixed point valuations of L_T over M.

The minimal fixed point. The set $\mathbf{FP}(L_T, M)$ can be equipped with a partial order $\leq$, by stipulating that:

$$V_M \leq V'_M \Leftrightarrow \forall \sigma \in Sen(L_T) : V_M(\sigma) = 1 \Rightarrow V'_M(\sigma) = 1$$

It can be shown (see Kripke [33]) that $(\mathbf{FP}_M(L_T), \leq)$ has a least element, V_M^{min}, which is called the *minimal fixed point valuation* over M. Kripke gives an intuitive motivation of V_M^{min} in terms of an imaginary subject that starts to assert (and deny) the truth-free sentences of L and, on the basis of these assertions, extends his assertoric actions to L_T sentences in a cumulative way. For instance, the imaginary subject asserts 'snow is white' and, on basis of that assertion, further asserts "snow is white' is true'. On the basis of that assertion, he then asserts, say, "snow is white' is true or grass is red' and, further, the truth of that disjunction. In a nutshell, Kripke's imaginary subject starts with assertoric actions that he is entitled to on the basis of the world (the ground model M) and he works his way *upwards*, by following the assertoric rules in an upwards

direction. Kripke shows that this upwards process culminates in the minimal fixed point V_M^{min}. It is in line with Kripke's *upwards story* of the imaginary subject to interpret the range of V_M^{min} as follows:

- $V_M^{min}(\sigma) = 1$: it is allowed to assert σ on basis of M.
- $V_M^{min}(\sigma) = \frac{1}{2}$: it is neither allowed to assert nor to deny σ on basis of M.
- $V_M^{min}(\sigma) = 0$: it is allowed to deny σ on basis of M.

Sentences which are valuated as $\frac{1}{2}$ by V_M^{min} will be called *ungrounded*. The Truthteller is, just like the Liar, an example of an ungrounded sentence. Below, we will consider an alternative interpretation of V_M^{min} that is backed up by a *downwards story*.

A SK fixed point valuation ensures the transparency of truth. That is, with ϕ' resulting from ϕ by replacing (one or more occurrences of) $T([\sigma])$ with σ and with V_M a SK fixed point valuation, we have that $V_M(\phi) = V_M(\phi')$. *Prima facie*, it seems that a SK fixed point valuation V_M ensures the transparency of truth via its second defining condition: the identity of truth. First looks are, in this case, deceiving, as we will now show. For, a *Supervaluation fixed point valuation* $S_M : Sen(L_T) \to \{0, \frac{1}{2}, 1\}$ also respects the ground model M and the identity of truth, but it violates the transparency of truth. To illustrate this, there's no need to define the notion of a Supervaluation fixed point valuation in detail; the following remarks suffice. Let S_M be a Supervaluation fixed point valuation and let $\neg T(\lambda)$ be a Liar. Then:

i. S_M valuates all classical tautologies as 1.

ii. $S_M(\sigma) = S_M(\sigma \vee \sigma)$, for all sentences σ of L_T.

iii. $S_M(\neg T(\lambda)) = \frac{1}{2}$.

From *i*, it follows that $S_M(\neg T(\lambda) \vee T(\lambda)) = 1$ and so, as λ denotes $\neg T(\lambda)$—and as we are allowed to substitute coreferential terms—we get that:

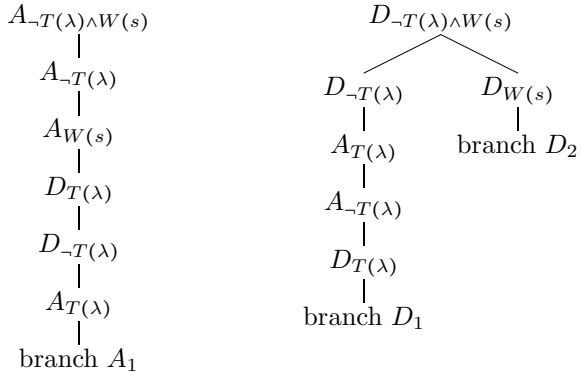
$$S_M(\neg T(\lambda) \vee T([\neg T(\lambda)])) = 1$$

From *ii* and *iii*, we know that $\neg T(\lambda) \vee \neg T(\lambda)$ is valuated as $\frac{1}{2}$ by S_M. But $\neg T(\lambda) \vee \neg T(\lambda)$ is obtained from $\neg T(\lambda) \vee T([\neg T(\lambda)])$ by substituting $T([\neg T(\lambda)])$ for $\neg T(\lambda)$. Hence, according to S_M "*x is true* and x are not intersubstitutable", i.e., S_M violates the transparency of truth.

(12) Two novel frameworks for theories of truth. It is now time to have a look at how, in this thesis, assertoric rules and assertoric norms are used to induce theories of truth which respect the transparency of truth. In this thesis we develop two novel—interrelated—approaches to induce such theories of truth: we develop two novel frameworks to define theories of truth. The frameworks are called *assertoric semantics* (developed in Section 4, see also Section 5 and Section 7) and the *method of closure games* (developed in Section 5, see also Section 6 and Section 7). Both frameworks define a theory of truth by specifying the assertoric status of L_T's sentences relative to an assertoric norm. In both frameworks, the assertoric status of a sentence σ is determined by considering (calculating) whether we can live up, according to the norm, to the

commitments that are associated with an assertion (denial) of σ. The frameworks differ in their use of the assertoric rules. Assertoric semantics may be interpreted as a *semantic version of* a signed tableau calculus for L_T, which (in an abstract sense) explains its use of the assertoric rules. The method of closure games, on the other hand, takes the assertoric rules (together with an assertoric norm) as constituting a two player *game*. Below, we turn to the two frameworks[13] in more detail.

(13) Assertoric semantics As we announced above, assertoric semantics is a *semantic version of* a signed tableau calculus for L_T: whereas tableau calculi define *proof systems*, assertoric semantics is a *semantic valuation tool*. Relative to a ground model M and upon specification of an assertoric norm ‡, assertoric semantics returns a semantic valuation $\mathcal{V}^{\ddagger}_{M}$ of the sentences of L_T. The valuation $\mathcal{V}^{\ddagger}_{M}$ tells us which sentences of L_T are assertible and / or deniable in M according to ‡. To obtain $\mathcal{V}^{\ddagger}_{M}$, assertoric semantics associates, with each sentence σ of L_T, two *assertoric trees*: $\mathfrak{T}^{\sigma}_{A}$ and $\mathfrak{T}^{\sigma}_{D}$. The assertion tree of σ, $\mathfrak{T}^{\sigma}_{A}$, keeps track of the commitments that are associated with an assertion of σ, whereas its denial tree $\mathfrak{T}^{\sigma}_{D}$ keeps track of the commitments that are associated with a denial of σ. Formally, an assertoric tree $\mathfrak{T}^{\sigma}_{X}$, where $X \in \{A, D\}$, is a set of branches and a branch of $\mathfrak{T}^{\sigma}_{X}$ is a *minimal* set of AD signed sentences, containing X_{σ} and *downwards saturated* under the assertoric rules. The following example, where $\neg T(\lambda)$ is the Liar and where $W(s)$ is an atomic sentence of L, suffices to get the idea:



So, the assertion tree of $\neg T(\lambda) \wedge W(s)$ contains a single branch, A_1, whereas its denial tree contains two branches: D_1 and D_2.

Whether or not one is able to live up to the commitments associated with an assertion (denial) of σ depends, in general, on the assertoric norm under consideration. Assertoric semantics formalizes assertoric norms as *closure conditions* on the branches of the assertoric trees. Closure conditions are necessary and sufficient conditions for a branch to be *closed*, whereas a branch that is not closed is called *open*. Intuitively, the closure of a branch indicates that it is not possible to live up to the assertoric commitments associated with that branch.

[13] In **(1)**, we announced that we develop *three* frameworks in which to define and study theories of truth. Besides assertoric semantics and the method of closure games, we also mentioned *the strict-tolerant calculus*. The strict-tolerant calculus, however, is not a framework to define theories of truth, but rather, to study the *consequence relations* induced by such theories as explained in **(18)** and **(19)** below.

An assertoric tree is called *closed* just in case al its branches are closed, and is called *open* otherwise. Intuitively, the closure of $\mathfrak{T}_A^\sigma$ under closure condition ‡ in M indicates that, relative to assertoric norm ‡, it is not possible to live up to the assertoric commitments associated with an assertion of σ in M; when $\mathfrak{T}_A^\sigma$ is *closed*$_‡$ *in* M we say that it is *forbidden* to assert σ according to ‡ in M. Dually, when $\mathfrak{T}_A^\sigma$ is *open*$_‡$ *in* M we say that it is *allowed* to assert σ according to ‡ in M. Hence, with ‡ an arbitrary closure condition and with M a ground model, assertoric semantics induces a theory of truth, $\mathcal{V}_M^‡$, as follows:

$$\mathcal{V}_M^‡(\sigma) = \begin{cases} \mathbf{a} := (1,0), & \mathfrak{T}_A^\sigma \text{ is open}_‡ \text{ in } M \ \ \& \ \mathfrak{T}_D^\sigma \text{ is closed}_‡ \text{ in } M \\ \mathbf{b} := (1,1), & \mathfrak{T}_A^\sigma \text{ is open}_‡ \text{ in } M \ \& \ \mathfrak{T}_D^\sigma \text{ is open}_‡ \text{ in } M \\ \mathbf{n} := (0,0), & \mathfrak{T}_A^\sigma \text{ is closed}_‡ \text{ in } M \ \& \ \mathfrak{T}_D^\sigma \text{ is closed}_‡ \text{ in } M \\ \mathbf{d} := (0,1), & \mathfrak{T}_A^\sigma \text{ is closed}_‡ \text{ in } M \ \& \ \mathfrak{T}_D^\sigma \text{ is open}_‡ \text{ in } M \end{cases}$$

The values **a**ssertible only, **b**oth assertible and deniable, **n**either assertible nor deniable and **d**eniable only, wear their interpretation on their sleeves.

Let us illustrate our abstract account of assertoric semantics via a concrete example. The *strict norm* as defined in (**7**) is, in fact, an example of a closure condition. Let us stipulate that $W(s)$ represents an attribution of whiteness to snow, in other words, the ground model M is such that $W(s)$ represents 'snow is white'. Let us see how $\mathcal{V}_M^s$, the assertoric valuation function induced by the strict norm, valuates $\neg T(\lambda) \wedge W(s)$. First, consider the assertion tree of $\neg T(\lambda) \wedge W(s)$. Note that its sole branch is closed, as it contains both an assertion and a denial of the Liar. Hence, the assertion tree of $\neg T(\lambda) \wedge W(s)$ is closed and so it is forbidden (according to the strict norm) to assert $\neg T(\lambda) \wedge W(s)$. Branch D_1 of the denial tree of $\neg T(\lambda) \wedge W(s)$ is closed for the same reason: it contains both an assertion and denial of the Liar. Branch D_2 of the denial tree is closed for a different reason: it contains a denial of 'snow is white'. As both its branches are closed, the denial tree of $\neg T(\lambda) \wedge W(s)$ is closed and so it is forbidden (according to the strict norm) to deny $\neg T(\lambda) \wedge W(s)$. Hence, $\neg T(\lambda) \wedge W(s)$ is *neither* assertible nor deniable according to the strict norm. The Truthteller $T(\tau)$ is an example of a sentence that is *both* assertible and deniable according to the strict norm, which is testified by the fact that:

$$\mathfrak{T}_A^{T(\tau)} = \{\{A_{T(\tau)}\}\} \qquad \mathfrak{T}_D^{T(\tau)} = \{\{D_{T(\tau)}\}\}$$

Note that, although $T(\tau)$ is both assertible and deniable according to the strict norm, it is not allowed to assert and deny the Truthteller "at the same time"[14]. Asserting and denying $T(\tau)$ "at the same time" is, according to assertoric semantics, tantamount to asserting $T(\tau) \wedge \neg T(\tau)$. And, as an inspection of the assertion tree of $T(\tau) \wedge \neg T(\tau)$ reveals, it is not allowed to assert $T(\tau) \wedge \neg T(\tau)$ according to the strict norm:

[14]Strictly speaking, it is *never possible* to assert and deny any sentence whatsoever at the same time: one (assertoric) action at a time. When we say that it is allowed to assert and deny σ "at the same time", we mean that it is allowed to take up a joint commitment to an assertion of σ and to a denial of σ. To take up such a commitment, we can assert $\sigma \wedge \neg\sigma$ (or deny $\sigma \vee \neg\sigma$).

$$A_{T(\tau)\wedge\neg T(\tau)}$$
$$|$$
$$A_{T(\tau)}$$
$$|$$
$$A_{\neg T(\tau)}$$
$$|$$
$$D_{T(\tau)}$$

More generally, it is, according to $\mathcal{V}^s_M$, *never* allowed to assert and deny a sentence "at the same time".

Although assertoric semantics is a novel framework to define and study theories of truth, $\mathcal{V}^s_M$ is a familiar *function*. For, $\mathcal{V}^s_M$ turns out to be identical to the function $\mathcal{K}^{\mathbf{4}}_M$, which Kripke [33] defined by quantifying over all SK fixed point valuations over M, i.e. by quantifying over $\mathbf{FP}(L_T, M)$:

- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = (1,0) \Leftrightarrow \exists V_M : V_M(\sigma) = 1$ and $\nexists V_M : V_M(\sigma) = 0$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = (1,1) \Leftrightarrow \exists V_M : V_M(\sigma) = 1$ and $\exists V_M : V_M(\sigma) = 0$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = (0,0) \Leftrightarrow \nexists V_M : V_M(\sigma) = 1$ and $\nexists V_M : V_M(\sigma) = 0$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = (0,1) \Leftrightarrow \nexists V_M : V_M(\sigma) = 1$ and $\exists V_M : V_M(\sigma) = 0$

For a proof that $\mathcal{V}^s_M = \mathcal{K}^{\mathbf{4}}_M$ and a discussion of the distinct philosophical interpretations that we associate with those functions, the reader is referred to Section 5. As $\mathcal{K}^{\mathbf{4}}_M$ is obtained by quantifying over all SK fixed point valuations over M and as a SK fixed point valuation satisfies the transparency of truth, it readily follows that $\mathcal{K}^{\mathbf{4}}_M$ satisfies the transparency of truth. Hence, as $\mathcal{V}^s_M = \mathcal{K}^{\mathbf{4}}_M$, $\mathcal{V}^s_M$ satisfies the transparency of truth.

Speaking of a *strict* norm suggests that we also have a *tolerant* norm around. Indeed we do. The tolerant norm is obtained by removing the third condition of the strict norm. That is, the closure conditions that represent the tolerant norm are as follows:

Tolerant norm A branch B is *closed* just in case B contains:
1) A_α, where α is an atomic sentence of L s.t. $\mathcal{C}_M(\alpha) = 0$, or
2) D_α, where α is an atomic sentence of L s.t. $\mathcal{C}_M(\alpha) = 1$.

Thus, in a catchy slogan, the tolerant norm may be stated as follows: *thou shalt respect the world*. We will use $\mathcal{V}^t_M$ to indicate the valuation function that is induced by the tolerant norm. In contrast to $\mathcal{V}^s_M$, the range of $\mathcal{V}^t_M$ consists of three values: **a**, **b** and **d**. Just like $\mathcal{V}^s_M$, the function $\mathcal{V}^t_M$ turns out to be a familiar one: modulo a translation of **a**, **b** and **d** as, respectively, $1, \frac{1}{2}$ and 0, $\mathcal{V}^t_M$ is identical to the SK minimal fixed point valuation V^{min}_M, as we prove in Section 7. Hence, $\mathcal{V}^t_M$ satisfies the transparency of truth.

Observe that $\mathcal{V}^t_M$'s assertoric interpretation of $V^{min}_M(\sigma) = \frac{1}{2}$ is diametrically opposed to the Kripkean interpretation that was outlined above in **(11)**. We will use $\mathcal{K}^{min}_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$ to denote the Kripkean interpretation of V^{min}_M, obtained by translating $1, \frac{1}{2}$ and 0 as, respectively, **a**, **n** and **d**. Indeed, according to $\mathcal{K}^{min}_M$, ungrounded sentences are *neither* assertible nor deniable, while they are *both* assertible and deniable according to $\mathcal{V}^t_M$.

Remember that, according to $\mathcal{V}^s_M$, sentences that are valuated as **b** are both

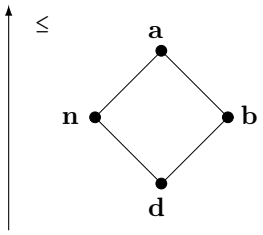
assertible and deniable, but *never* "at the same time". According to $\mathcal{V}^t_M$ however, sentences that are valuated as **b** are *always* both assertible and deniable "at the same time". The last remark readily follows from the Strong Kleene compositionality of $\mathcal{V}^t_M$: when $\mathcal{V}^t_M(\sigma) = \mathbf{b}$, we get that $\mathcal{V}^t_M(\neg\sigma) = \mathbf{b}$, and so $\mathcal{V}^t_M(\sigma \wedge \neg\sigma) = \mathbf{b}$, i.e., it is allowed to assert (and deny) $\sigma \wedge \neg\sigma$.

For sure, the fact that sentences of form $\sigma \wedge \neg\sigma$ may be (tolerantly) assertible is counterintuitive. According to $\mathcal{K}^{min}_M$, sentences of form $\sigma \wedge \neg\sigma$ are never assertible. In light of this observation, it is tempting to say that $\mathcal{K}^{min}_M$ is a more natural interpretation of V^{min}_M than $\mathcal{V}^t_M$. On the other hand, the intuition that it is never allowed to assert a sentence of form $\sigma \wedge \neg\sigma$ seems to be closely related to the intuition that it is always allowed to deny a sentence of form $\sigma \wedge \neg\sigma$. Focussing on the latter intuition, the roles are exactly reversed: according to $\mathcal{V}^t_M$ it is always allowed to deny a sentence of form $\sigma \wedge \neg\sigma$, whereas this is not the case according to $\mathcal{K}^{min}_M$. More generally, it seems that there is an important sense in which the vices and virtues of $\mathcal{K}^{min}_M$ are on a par with, respectively, the virtues and vices of $\mathcal{V}^t_M$. However, from the perspective of assertoric semantics, there is an important distinction between $\mathcal{V}^t_M$ and $\mathcal{K}^{min}_M$. For, according to assertoric semantics, a sentence is valuated as **b** just in case both its assertoric trees are *open*, whereas a sentence is valuated as **n** just in case both its assertoric trees are *closed*. We defined $\mathcal{V}^t_M$ via the tolerant closure conditions, according to which both assertoric trees of ungrounded sentences are open. On the other hand, I do not know how to pick closure conditions that induce $\mathcal{K}^{min}_M$ and according to which, in particular, both assertoric trees of ungrounded sentences are closed. Thus, in contrast to $\mathcal{K}^{min}_M$, $\mathcal{V}^t_M$ has a natural definition in the framework of assertoric semantics.

(14) $\mathcal{V}^s_M$, the entitlement reading and compositionality According to the *entitlement reading* (cf.**(6)**) of the assertoric rules, the upwards direction of the $A_\wedge$ rule reads as follows:

> it is allowed to assert α and it is allowed to assert $\beta \Rightarrow$
> it is allowed to assert $\alpha \wedge \beta$.

Above, we explained that according to $\mathcal{V}^s_M$, it is allowed to assert the Truthteller $T(\tau)$. Similarly, it is allowed to assert the negation of the Truthteller, $\neg T(\tau)$, according to $\mathcal{V}^s_M$. However, we also explained that it is *not* allowed to assert $T(\tau) \wedge \neg T(\tau)$ according to $\mathcal{V}^s_M$. Hence, $\alpha := T(\tau)$ and $\beta := \neg T(\tau)$ testify that the upwards direction of the entitlement reading of the $A_\wedge$ rule is not valid according to $\mathcal{V}^s_M$. In Section 4, we show that $\mathcal{V}^s_M$ validates the (entitlement reading of) the assertoric rules in downwards direction though. Further, we show that an arbitrary 4-valued assertoric valuation function $V_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{n}, \mathbf{b}, \mathbf{d}\}$ validates the assertoric rules in both directions just in case V_M respects the identity of truth and defines a *4-valued Strong Kleene semantics*, a notion that is conveniently explained via the Hasse diagram of $FOUR = (\{\mathbf{a}, \mathbf{n}, \mathbf{b}, \mathbf{d}\}, \leq)$:



V_M is said to define a *4-valued Strong Kleene semantics* just in case: 1) $\neg$ swaps **a** for **d** and vice versa, while it acts as the identity operation on **b** and **n**, 2) $\wedge$ and $\vee$ act as meet and join in *FOUR*, and 3) $\exists$ and $\forall$ behave as generalized $\vee$ and $\wedge$ respectively.

Hence, as $\mathcal{V}^s_M$ does not validate the assertoric rules in both directions, it does not define a 4-valued Strong Kleene semantics. Indeed, $\mathcal{V}^s_M$ does not define a *compositional* semantics, in the sense that the semantic value of a complex sentence can not be explained in terms of the semantic values of its constituents. Compositionality is often cited of as an attractive property of a semantics:

> Proponents of compositionality typically emphasize the productivity and systematicity of our linguistic understanding. We can understand a large—perhaps infinitely large—collection of complex expressions the first time we encounter them, and if we understand some complex expressions we tend to understand others that can be obtained by recombining their constituents. Compositionality is supposed to feature in the best explanation of these phenomena. (Szabó, [51])

Although $\mathcal{V}^s_M$ does not define a compositional semantics in the sense alluded to above, I'm not sure whether this implies that it is not compositional in the sense of Szabó's remark. For instance, there is a sense in which $\mathcal{V}^s_M$'s semantic valuation of $\alpha \wedge \beta$ (via the assertoric trees of $\alpha \wedge \beta$) can be understood as a "recombination" of the assertoric trees of α and β. We will not enter this broader discussion on ($\mathcal{V}^s_M$'s) compositionality. In what follows, we are only concerned with the notion of compositionality according to which $\mathcal{V}^s_M$'s treatment of the Truthteller, its negation and their conjunction testifies that $\mathcal{V}^s_M$ *is not compositional*. According to this notion of compositionality, $\mathcal{V}^t_M$ *is compositional*.

There are many senses in which one can answer the following question: *why* is $\mathcal{V}^s_M$ not compositional? For instance, one can cite the definition of compositionality and show that $\mathcal{V}^s_M$ does not satisfy it; indeed, in that sense we already answered the question. Below, we will be concerned with the question '*why* is $\mathcal{V}^s_M$ not compositional?' in the following sense:

> The valuations $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$ are closely related, in the sense that they arise out of the same assertoric rules and out of related closure conditions. Why then, is $\mathcal{V}^t_M$ compositional while $\mathcal{V}^s_M$ is not? Can we, more generally, understand the (non-) compositionality of a valuation function in terms of the *closure conditions* that induce it?

In Section 5 we answer the last question affirmatively. There, we show how to *characterize* (3- and 4- valued) Strong Kleene compositionality in terms of closure conditions. However, the closure conditions in Section 5 are not defined over *branches* of *assertoric trees*, but rather over *expansions* that are induced in *closure games*. That is, in Section 5 we develop *the method of closure games*. As a semantic valuation method, the method of closure games is best understood as a *refinement* of assertoric semantics. Whereas a branch is a *set* of AD sentences, an expansion is an (infinite) *sequence* of AD sentences. An expansion of X_σ can be thought of as "running through" (or "lying in") a branch of $\mathfrak{T}^\sigma_X$. Hence, we may say that the notion of an expansion is more fine-grained than the notion of a branch. Below, we explain how, by putting closure conditions on expansions, the method of closure games induces valuations of L_T.

(15) The Method of Closure Games In a *closure game*, there are two players, called $\sqcup$ and $\sqcap$. Player $\sqcup$ controls all AD sentences of *disjunctive type* and player $\sqcap$ controls all sentences of *conjunctive type*. Sentences of form $A_{\alpha\vee\beta}, D_{\alpha\wedge\beta}, A_{\exists\phi(x)}, D_{\forall\phi(x)}$ are of disjunctive type, all others are of conjunctive type. A *strategy* of a player is a mapping of each AD sentence X_σ that is in his control to exactly one of the *immediate successors* of X_σ, as specified by the assertoric rule applicable to X_σ. A few examples suffice to illustrate the notion of a strategy. The immediate successors of $A_{\alpha\wedge\beta}$ are A_α and A_β and, as $A_{\alpha\wedge\beta}$ is of conjunctive type, a strategy of player $\sqcap$ maps $A_{\alpha\wedge\beta}$ to either A_α or A_β. As $A_{T(\overline{\sigma})}$ has only one immediate successor, A_σ, every strategy of player $\sqcap$ must map $A_{T(\overline{\sigma})}$ to A_σ. A strategy for player $\sqcup$, who controls $D_{\alpha\wedge\beta}$, maps $D_{\alpha\wedge\beta}$ to either D_α or D_β.

With f a strategy for player $\sqcup$, g a strategy for player $\sqcap$ and with X_σ an arbitrary AD sentence, the tuple (X_σ, f, g) defines an *expansion of* X_σ. In general, an expansion of X_σ is an infinite[15] sequence of AD sentences whose first element is X_σ and whose successor relation respects the assertoric rules. As an example, here is the expansion of $A_{\neg T(\lambda)}$, i.e., of an assertion of the Liar:

$$A_{\neg T(\lambda)}, D_{T(\lambda)}, D_{\neg T(\lambda)}, A_{T(\lambda)}, A_{\neg T(\lambda)} \ldots \tag{1.10}$$

Indeed, $A_{\neg T(\lambda)}$ has only one expansion and so, in *the closure game for* $A_{\neg T(\lambda)}$, none of the players can influence the expansion of $A_{\neg T(\lambda)}$ that is realized. In general, an AD sentence X_σ may have (infinitely) many expansions, each of which is realized by some strategy pair (f, g) of our players. For instance, $A_{P(c_1)\wedge P(c_2)}$, where $P(c_1)$ and $P(c_2)$ are atomic sentences of L, has two expansions and, in *the closure game for* $A_{P(c_1)\wedge P(c_2)}$, player $\sqcap$ can determine which one is realized. By setting $g(A_{P(c_1)\wedge P(c_2)}) = A_{P(c_1)}$, player $\sqcap$ ensures that expansion (1.11) is realized, while $g(A_{P(c_1)\wedge P(c_2)}) = A_{P(c_2)}$ realizes expansion (1.12).

$$A_{P(c_1)\wedge P(c_2)}, A_{P(c_1)}, A_{P(c_1)}, A_{P(c_1)}, \ldots \tag{1.11}$$

$$A_{P(c_1)\wedge P(c_2)}, A_{P(c_2)}, A_{P(c_2)}, A_{P(c_2)}, \ldots \tag{1.12}$$

We will write $\mathsf{exp}(X_\sigma, f, g)$ to denote the expansion of X_σ that is induced by strategies f (for player $\sqcup$) and g (for player $\sqcap$).

A *closure function* † assigns, to each ground model M, a *closure condition*

[15] Whenever an expansion "hits" a signed atomic sentence of L it keeps on repeating it indefinitely.

$\dagger(M) = \{O^\dagger_M, C^\dagger_M\}$, consisting of the sets $O^\dagger_M$ and $C^\dagger_M$ of all open and all closed expansions in M. In a closure game for X_σ played in M relative to $\dagger$, player $\sqcup$ tries to pick his strategy f in such a way that the expansion of X_σ that is realized will be contained in $O^\dagger_M$. We will write $O^\dagger_M(X_\sigma)$, and say that X_σ *is open relative to* $\dagger(M)$, to indicate that player $\sqcup$ has a strategy which *ensures* that the expansion of X_σ ends up in $O^\dagger_M$. That is:

$$O^\dagger_M(X_\sigma) \Leftrightarrow \exists f \forall g \;\; \mathsf{exp}(X_\sigma, f, g) \in O^\dagger_M \tag{1.13}$$

X_σ is *closed relative to* $\dagger(M)$, denoted $C^\dagger_M(X_\sigma)$, just in case not $O^\dagger_M(X_\sigma)$. As specified by (1.13), a closure condition for expansions induces a closure condition for AD sentences. The closure condition for AD sentences is used to induce a valuation for L_T, denoted $\mathcal{V}^\dagger_M$:

$$\mathcal{V}^\dagger_M(\sigma) = \begin{cases} \mathbf{a} := (1,0), & O^\dagger_M(A_\sigma) \text{ and } C^\dagger_M(D_\sigma); \\ \mathbf{b} := (1,1), & O^\dagger_M(A_\sigma) \text{ and } O^\dagger_M(D_\sigma); \\ \mathbf{n} := (0,0), & C^\dagger_M(A_\sigma) \text{ and } C^\dagger_M(D_\sigma); \\ \mathbf{d} := (0,1), & C^\dagger_M(A_\sigma) \text{ and } O^\dagger_M(D_\sigma). \end{cases}$$

In general $\mathcal{V}^\dagger_M$ may, but need not have, a range of four values. The intuitive interpretation of the functions that are induced by the method of closure games resembles that of assertoric semantics. For instance, $\mathcal{V}^\dagger_M(\sigma) = \mathbf{a}$ indicates that it is allowed to assert, but not to deny, sentence σ in ground model M according to the norms for assertion and denial that are specified by $\dagger$. In a sense, assertoric semantics and the method of closure games are two distinct ways of formalizing a single intuition. Here, we will not enter the question as to whether assertoric norms are "better" modeled via closure conditions on branches or on expansions: for some remarks pertaining to that question, see Section 5.7. Rather, we will sketch how the method of closure games allows us to characterize all 3- and 4-valued Strong Kleene fixed point valuations in a uniform manner. In particular, we will sketch how the method of closure games allows us to characterize (3- and 4- valued) Strong Kleene compositionality in terms of closure conditions.

For each expansion exp, its *successor expansion* exp', is obtained by deleting the first term of exp. For instance, (1.14) is the successor expansion of (1.11).

$$A_{P(c_1)}, A_{P(c_1)}, A_{P(c_1)}, \ldots \tag{1.14}$$

A closure condition $\dagger(M) = \{O^\dagger_M, C^\dagger_M\}$ satisfies the *Stable Judgement Constraint* (SJC), just in case for every expansion exp we have that:

$$\mathsf{SJC}: \qquad \mathsf{exp} \in C^\dagger_M \Leftrightarrow \mathsf{exp}' \in C^\dagger_M$$

If a closure condition $\dagger(M)$ satisfies SJC, the judgement of $\dagger$ as to whether an expansion is open or closed is stable, in the sense that it does not change along the expansion. The content of our *first stable judgement theorem* (See Section 5) is that closure conditions which satisfy SJC induce Strong Kleene valuations:

First stable judgement theorem
Let M be a ground model, let $\dagger(M)$ be a closure condition which satisfies SJC and let $\mathcal{V}^\dagger_M$ be the valuation function induced by the method of closure games. Then:

1. $\mathcal{V}^{\dagger}_{M}$ has either a classical, a 3- or 4- valued Strong Kleene semantics.

2. Whenever $\overline{\sigma}$ denotes σ in M: $\mathcal{V}^{\dagger}_{M}(T(\overline{\sigma})) = \mathcal{V}^{\dagger}_{M}(\sigma)$. That is, $\mathcal{V}^{\dagger}_{M}$ respects the identity of truth.

Note that the first stable judgement theorem is one sided: *if* closure conditions satisfies SJC *then* they induce a Strong Kleene valuation[16] which respects the identity of truth. The converse direction does not hold: in Section 5 we show that there are closure conditions which violate SJC and which, nevertheless, induce a Strong Kleene valuation function which respects the identity of truth. However, *the second stable judgement theorem* (see Section 5) comes close to a converse reading of the first stable judgement theorem.

Second stable judgement theorem
Let M be a ground model and let V_M be a 2, 3 or 4 valued Strong Kleene valuation of L_T in M which respects the identity of truth. Then there is a closure condition $\dagger(M)$ which respects SJC and such that, with $\mathcal{V}^{\dagger}_{M}$ the valuation induced by $\dagger(M)$, we have that: $\mathcal{V}^{\dagger}_{M} = V_M$.

Remember that, in **(14)**, we raised the following question: *Can we, more generally, understand the (non-) compositionality of a valuation function in terms of the closure conditions that induce it?* Together, the first and second stable judgement theorem testify that we can answer with a well-known campaign slogan: *yes, we can.*

Observe that a closure condition $\dagger(M)$ which satisfies SJC gives rise to a Strong Kleene valuation function $\mathcal{V}^{\dagger}_{M}$ which respects the identity of truth, but also, that $\mathcal{V}^{\dagger}_{M}$ need not respect the ground model M. As such, there is no guarantee that $\mathcal{V}^{\dagger}_{M}$ is a (2-, 3- or 4- valued) Strong Kleene *fixed point valuation over* M. To ensure that we induce a Strong Kleene fixed point valuation over M, we need to impose a further (obvious) constraint on our closure conditions. The (world-respecting) constraint is formulated in terms of *grounded* expansions. We say that an expansion is *grounded* just in case it hits a signed atomic sentence of L and *ungrounded* otherwise. Thus, expansions (1.11) and (1.12) are grounded, whereas (1.10) is ungrounded. We say that an expansion exp is *grounded and correct in* M just in case exp is grounded and, with X_σ the (unique) signed atomic sentence of L that occurs on exp, we have that:

- $(X_\sigma = A_\sigma$ and $\mathcal{C}_M(\sigma) = 1)$ or $(X_\sigma = D_\sigma$ and $\mathcal{C}_M(\sigma) = 0)$.

In Section 5, we show that closure conditions $\dagger(M)$ which satisfy SJC and the *world respecting constraint* (WRC), induce Strong Kleene fixed point valuations, where:

$$\mathsf{WRC}: \begin{cases} \{\mathsf{exp} \mid \mathsf{exp} \text{ is grounded and correct in } M\} \subseteq O^{\dagger}_{M}, \text{ and} \\ \{\mathsf{exp} \mid \mathsf{exp} \text{ is grounded and incorrect in } M\} \subseteq C^{\dagger}_{M}. \end{cases}$$

Further, we show that the following closure conditions induce the Kripkean interpretation of the minimal fixed point valuation, i.e., $\mathcal{K}^{min}_{M} : Sen(L_T) \to \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$:

exp is closed in $M \Leftrightarrow$ exp is ungrounded or grounded and incorrect in M

[16] We may understand a classical valuation as a 2-valued Strong Kleene valuation.

So, although I do not know how to define $\mathcal{K}_M^{min}$ in assertoric semantics, $\mathcal{K}_M^{min}$ has a very natural definition in the method of closure games. In Section 7, we show that the counterpart of $\mathcal{K}_M^{min}$, the tolerant valuation $\mathcal{V}_M^t : Sen(L_T) \to \{\mathbf{a}, \mathbf{b}, \mathbf{d}\}$, can also be defined in the method of closure games; $\mathcal{V}_M^t$ is induced by the following closure conditions.

$$\mathsf{exp} \text{ is closed in } M \Leftrightarrow \mathsf{exp} \text{ is grounded and incorrect in } M \tag{1.15}$$

The fact that an expansion of X_σ can be thought of as "running through" a branch of $\mathfrak{T}_X^\sigma$ ensures that there are many interesting connections between assertoric semantics and the method of closure games. An example of such a connection is the definability of $\mathcal{V}_M^t$ in both frameworks via related closure conditions. For more examples, the reader is referred to Section 5. Most notably, we show (Section 5.5) how the *maximal intrinsic fixed point* can be defined by combining both frameworks.

(16) Generalized Strong Kleene fixed point valuations So, as a consequence of our first and second stable judgement theorem, the method of closure games is a powerful tool to study three and four valued SK fixed point valuations *in a uniform manner*. The uniform approach allows us to combine suitably related three and four valued SK fixed point valuations into *Generalized Strong Kleene fixed point valuations* (GSK fixed point valuations), a notion that is novel to this thesis. Let us illustrate the notion of a GSK fixed point valuation by means of $\mathbb{V}^{\mathbf{8}^+}$, an 8-valued GSK valuation which is defined in terms of one 4-valued and two 3-valued Strong Kleene theories.

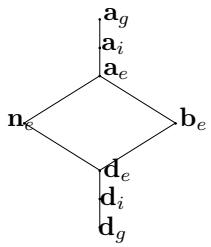


Figure 1.1: Hasse diagram of $\mathbf{8}^+_\leq$, the lattice of $\mathbb{V}^{\mathbf{8}^+}$.

The GSK compositionality of $\mathbb{V}^{\mathbf{8}^+}$ is explained as follows. Conjunction and disjunction behave as meet and join in the lattice $\mathbf{8}^+_\leq$, while universal and existential generalization behave as generalized meet and join. The only difference with SK compositionality concerns negation, which interchanges, for $x \in \{g, i, e\}$, $\mathbf{a}_x$ with $\mathbf{d}_x$ (and vice versa) and acts as the identity on $\mathbf{n}_e$ and $\mathbf{b}_e$. As explained in Section 5, we can think of $\mathbb{V}^{\mathbf{8}^+}$ as combining three assertoric norms, associated with the subscripts g, i and e that flank the assertoric values $\mathbf{a}, \mathbf{b}, \mathbf{n}$ and $\mathbf{d}$. More generally, Section 5 exploits the method of closure games to study conditions under which SK fixed point valuations can be combined into GSK ones.

(17) Desiderata for truth and GSK fixed point valuations In Section 6, we exploit the notion of a GSK fixed point valuation to argue that the *Modified Gupta-Belnap Desideratum* (**MGBD**), which is a desideratum for theories of truth due to Philip Kremer [32], has to be reconsidered. Intuitively, **MGBD**

says that if there is no vicious reference according to a theory of truth **T** (a formally defined notion that we will not discuss here) then, according to **T**, truth should behave like a classical concept (another formally defined notion that we will not discuss here). Formally:

MGBD If **T** *dictates that there is no vicious reference in ground model* M then **T** *dictates that truth behaves like a classical concept in ground model* M.

With respect to the *rationale* of **MGBD**, Kremer cites Gupta [23]:

> For models M belonging to a certain class—a class that we have not formally defined but which in intuitive terms contains models that permit only benign kinds of self-reference—the theory should entail that all Tarski biconditionals are assertible in the model M. (Gupta, [23, p19]

Thus, the proposed rationale for **MGBD** is that it is a theory-relative formalization of an intuitive desideratum that was formulated by Gupta. In [60] an **A**lternative—to **MGBD**—formalization of Gupta's **D**esideratum is proposed:

AD If **T** *dictates that there is no vicious reference in ground model* M then **T** *dictates that all the Tarski biconditionals are strongly assertible in* M.

In Section 6, we argue that **AD** is preferable over **MGBD** as a desideratum for theories of truth. For one thing, it seems to be superior to **MGBD** in capturing the rationale that is given for that desideratum. Further, we show that any theory of truth which violates **AD** violates **MGBD**, but also that there are theories of truth with a GSK semantics which violate **MGBD** while they satisfy **AD**. I take it that these results testify that the notion of a GSK fixed point valuation is a philosophically fruitful notion.

(18) The assertoric conception of truth, inferentialism and the strict-tolerant calculus. According to the assertoric conception of truth (see **(8)**) that is laid out in this thesis, the meaning of truth is given by the assertoric rules of the truth predicate. The assertoric conception of truth is naturally wedded to a broader account of meaning: that of *inferentialism*[17]. In a nutshell, inferentialism is the view that meanings are to be explained in terms of *which inferences are valid*. As assertoric semantics and the method of closure games do not inform us about which inferences are valid, there is a sense in which the assertoric conception of truth as articulated thus far is incomplete. To fill in the lacuna, Section 7 presents *the strict-tolerant calculus*, which can be thought of as *the syntactic counterpart of assertoric semantics*. As we will explain below (cf. **(19)**), the strict-tolerant calculus is a signed tableau system which gives us syntactic characterizations of *four fixed point consequence relations* that are semantically defined by quantifying over the class of all (3-valued Strong Kleene) fixed point valuations. To develop the strict tolerant calculus, we follow Cobreros et al. [11], in drawing a distinction between *strict* assertions and denials

[17] We thus develop a deflationary, assertoric conception of truth which is naturally wedded to inferentialism. As such, the conception of truth is close in spirt to Horsten's *inferential deflationism*, a position that he develops in [27]. For some remarks on the relation between assertoric semantics and Horsten's inferential deflationism, see Section 4.

and *tolerant* assertions and denials. We exploit the distinction by equipping our calculus with four signs, which are naturally associated with the four different speech acts that arise out of the strict-tolerant distinction. As such, the strict-tolerant calculus does not single out a particular fixed point consequence relation as favorite. However, the strict-tolerant calculus, in combination with assertoric semantics, sheds light on the *Strict Tolerant Conception of Truth* (STCT), a recent conception of truth that is articulated by Dave Ripley [46]. An essential ingredient of STCT is its commitment to a particular fixed point consequence relation. Below, we first present the rationale of the strict-tolerant calculus (cf. **(19)**), after which we explain the relation between the strict-tolerant calculus and assertoric semantics (cf. **(20)**). Then, we explain how the strict-tolerant calculus and assertoric semantics jointly shed light on STCT (cf. **(21)**).

(19) The strict-tolerant calculus In our presentation of assertoric semantics, we discussed two theories of truth, $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$. We presented $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$ as describing which assertions and denials are allowed according to, respectively, the *tolerant norm* and the *strict norm*. This presentation suggests the following broader picture: there are *two* assertoric speech acts, assertion and denial, while the strict norm and the tolerant norm are *two* assertoric norms that (may) govern our assertoric practice. In what follows we reverse this picture, in accordance with the strict-tolerant distinction, as follows: there are *four* assertoric speech acts, strict assertion, strict denial, tolerant assertion and tolerant denial, while there is a *single* assertoric norm: the strict-tolerant one. As we will explain below, in a sense, the strict-tolerant calculus *forces us* to switch to the revised picture.

Let us first explain how the strict-tolerant slang is to be used with respect to a (3-valued Strong Kleene) fixed point valuation $V : Sen(L_T) \to \{0, \frac{1}{2}, 1\}$. Sentences that are valuated as 1 are *strictly assertible*, sentences that are valuated as 0 are *strictly deniable* and sentences that are valuated as $\frac{1}{2}$ are *neither* strictly assertible nor strictly deniable. Sentences that receive a value in $\{1, \frac{1}{2}\}$ are *tolerantly assertible*, whereas those that receive a value in $\{0, \frac{1}{2}\}$ are *tolerantly deniable*. Indeed, sentences that are valuated as $\frac{1}{2}$ are neither strictly assertible nor deniable, but, at the same time, both tolerantly assertible and deniable. Exploiting the strict-tolerant terminology, we can define the four *fixed point consequence relations*, $\vDash^{ss}$, $\vDash^{tt}$, $\vDash^{st}$ and $\vDash^{ts}$ where, with $i, j \in \{t, s\}$, $\Gamma \vDash^{ij} \Delta$ means that: whenever all premisses in Γ are *i-ly* assertible, some conclusion in Δ is *j-ly* assertible. For instance:

$\Gamma \vDash^{st} \Delta$ iff for every fixed point valuation V (over *some* ground model) :

$$\forall \alpha \in \Gamma : V(\alpha) = 1 \Rightarrow \exists \beta \in \Delta : V(\beta) \in \{1, \tfrac{1}{2}\}$$

"All premisses *s*trictly assertible ⇒ some conclusion *t*olerantly assertible".

The strict-tolerant calculus is a signed tableau calculus which can be used to characterize $\vDash^{ss}$, $\vDash^{tt}$, $\vDash^{st}$ and $\vDash^{ts}$ in a uniform manner. The signs that are employed by the strict-tolerant calculus indicate the strict and tolerant assertoric actions: A^s, D^s, A^t and D^t indicate, respectively, a strict assertion, a strict denial, a tolerant assertion and a tolerant denial. A tableau calculus consists of (tableau expansion) *rules* and *closure conditions*, which specify when a tableau

path is closed. The rules of the strict-tolerant calculus are simply obtained by *doubling the assertoric rules*: each assertoric rule gives rise to a strict version and a tolerant one. In **(6)**, we displayed the assertoric rules for the truth predicate, negation and conjunction. With $i \in \{t, s\}$, these rules give rise to the following rules of the strict-tolerant calculus:

$$\frac{A^i_{T([\sigma])}}{A^i_\sigma} \qquad \frac{D^i_{T([\sigma])}}{D^i_\sigma} \qquad\qquad \frac{A^i_{\neg\alpha}}{D^i_\alpha} \qquad \frac{D^i_{\neg\alpha}}{A^i_\alpha} \qquad\qquad \frac{A^i_{\alpha\wedge\beta}}{A^i_\alpha, A^i_\beta} \qquad \frac{D^i_{\alpha\wedge\beta}}{D^i_\alpha \mid D^i_\beta}$$

So, the strict-tolerant calculus clearly points out that the rules that govern our strict assertoric actions are the same as the rules that govern our tolerant ones. Still, a sentence may be tolerantly assertible (deniable) without being strictly assertible or deniable, the Liar being a case in point. This distinction is not explained by the rules of our calculus, but rather by its closure conditions. Here are the closure conditions:

Closure conditions of the strict-tolerant calculus A tableau is closed just in case all its paths are closed. A path $\mathcal{P}$ of a tableau is closed just in case one of the following four conditions holds:

1. For some sentence σ of L_T: $\{A^s_\sigma, D^s_\sigma\} \subseteq \mathcal{P}$
2. For some sentence σ of L: $\{A^t_\sigma, D^t_\sigma\} \subseteq \mathcal{P}$
3. For some sentence σ of L_T: $\{A^s_\sigma, D^t_\sigma\} \subseteq \mathcal{P}$
4. For some sentence σ of L_T: $\{A^t_\sigma, D^s_\sigma\} \subseteq \mathcal{P}$

The closure conditions have the following rationale: 1) it is forbidden to[18] strictly assert and deny the same sentence. 2) it is forbidden to tolerantly assert and deny the same sentence of L. 3) it is forbidden to strictly assert and tolerantly deny the same sentence. 4) it is forbidden to tolerantly assert and strictly deny the same sentence. The closure conditions of the strict-tolerant calculus can be thought of as representing the *strict-tolerant norm* that governs the strict and tolerant assertoric actions. As reflected by closure conditions 3 and 4, the strict-tolerant norm is more than the sum of the strict norm and the tolerant norm. Above, we asserted that 'in a sense, the strict-tolerant calculus *forces us* to switch to the revised picture', i.e., to a picture with four speech acts and one norm. The ground for that assertion is the fact that the strict-tolerant norm is more than the sum of the strict norm and the tolerant norm.

A notion that plays a crucial role in (tableau based) soundness and completeness proofs for classical logic is that of *satisfiability*. In our soundness and completeness proofs pertaining to the fixed point consequence relations, a similar role is played by the notion of *fixed point satisfiability*. A set of (strict-tolerant) assertoric sentences S is said to be *fixed point satisfiable* just in case there exists a fixed point valuation V such that:

$$A^s_\sigma \in S \Rightarrow V(\sigma) = 1, \qquad D^s_\sigma \in S \Rightarrow V(\sigma) = 0, \tag{1.16}$$

[18] More precisely, it is forbidden to perform assertoric actions which enforce you to take up a commitment to a strict assertion and strict denial of the same sentence. We will not be that precise though.

$$A^t_\sigma \in S \Rightarrow V(\sigma) \in \{1, \tfrac{1}{2}\}, \qquad D^t_\sigma \in S \Rightarrow V(\sigma) \in \{0, \tfrac{1}{2}\}. \tag{1.17}$$

The notion of fixed point satisfiability allows us to reformulate the fixed point consequence relations in terms of sets of assertoric sentences. For instance, observe that per definition:

$$\Gamma \models^{st} \Delta \Leftrightarrow \{A^s_\alpha \mid \alpha \in \Gamma\} \cup \{D^s_\beta \mid \beta \in \Delta\} \text{ is not fixed point satisfiable.}$$

The reformulation of the fixed point consequence relations in terms of assertoric sentences motivates our definition of $\vdash^{st}$, $\vdash^{ss}$, $\vdash^{tt}$ and $\vdash^{ts}$. For instance, we have that:

$$\begin{aligned}\Gamma \vdash^{st} \Delta \Leftrightarrow &\text{ there exists a tableau starting with} \\ &\{A^s_\alpha \mid \alpha \in \Gamma\} \cup \{D^s_\beta \mid \beta \in \Delta\} \text{ that is closed.}\end{aligned}$$

In Section 7, we prove that $\Gamma \vdash^{st} \Delta$ just in case $\Gamma \models^{st} \Delta$. More generally, we prove the following theorem.

Strict-tolerant theorem: With $i, j \in \{s, t\}$: $\Gamma \vdash^{ij} \Delta \Leftrightarrow \Gamma \models^{ij} \Delta$

While the first and second stable judgement testify that the method of closure games gives us a uniform approach to all SK fixed point valuations, the strict-tolerant theorem testifies that the strict-tolerant calculus gives us a uniform approach to the SK fixed point consequence relations.

(20) The strict-tolerant calculus and assertoric semantics Assertoric semantics can be thought of as the semantic version of the strict-tolerant calculus. Dually, the strict-tolerant calculus can be thought of as the syntactic version of assertoric semantics. The connection between assertoric semantics and the strict tolerant calculus is nicely illustrated via $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$. Observe that the functions $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$ (which, in assertoric semantics, are thought of as deriving from the strict norm and tolerant norm respectively) can be induced by the closure conditions of the strict-tolerant calculus (modeling the *strict-tolerant norm*), augmented with the closure conditions associated with M. The function $\mathcal{V}^s_M$ of assertoric semantics can be thought of as answering the question which sentences are *initially*, strictly assertible and deniable, while $\mathcal{V}^t_M$ answers the same question in tolerant terms. On the other hand, by performing assertoric actions we take up certain assertoric commitments, which (may) rule out certain other assertoric actions as forbidden. For instance, strictly asserting a sentence rules out tolerantly denying it. More generally, the transmission of assertoric entitlements due to (strict and tolerant) assertions and denials is captured by the strict-tolerant calculus. For instance, we have that:

$\Gamma \vdash^{st} \Delta \Leftrightarrow$ strictly asserting all of Γ *rules out* strictly denying all of Δ.

Let's call the view according to which $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$ describe the initial assertoric entitlements and according to which the strict-tolerant calculus describes the transmission of assertoric entitlements the *Dynamic Conception of Strict-Tolerant assertion and denial* (DCST). According to DCST, the "rules out interpretation" of the four $\vdash^{ij}$ relations as illustrated for $\vdash^{st}$ above, are all *truisms* about the strict and tolerant assertoric actions. DCST doesn't single out one

of those relations as privileged, but treats them on a par. In contrast, STCT singles out $\vdash^{st}$ as privileged. However, STCT's judgment that $\vdash^{st}$ is privileged depends on its interpretation of the $\vdash^{ij}$ relations as *consequence relations*. Upon being interpreted as consequence relations, the four relations are certainly not on a par, as we explain below.

(21) The Strict Tolerant Conception of Truth. In [46], Ripley advocates a new approach to truth and semantics, which heavily relies on the strict-tolerant distinction. Ripley's conception of truth will be called the *Strict Tolerant Conception of Truth*. Most notably, STCT relies on an inferentialist theory of meaning according to which (a syntactic characterization of) $\vDash^{st}$ is the norm according to which inferences should be valuated as (in)correct. The reason that, according to STCT, inferences should be judged by the standards of $\vDash^{st}$ is, in a nutshell, that it is *well-behaved as a consequence relation*. In particular, STCT claims that it is better behaved as a consequence relation than the other three fixed point consequence relations.

Let us first illustrate the distinction between the "rules out interpretation" of $\vDash^{ij}$ and its "consequence interpretation". To do so, consider $\vDash^{ts}$ first. Importantly, it does *not* hold that:

$$\sigma \vDash^{ts} \sigma \tag{1.18}$$

On the "rules out interpretation" of $\vDash^{ts}$, it is very natural that (1.18) does not hold. For, tolerantly asserting a Truthteller or Liar does not rule out that you also deny it tolerantly. However, on the "consequence interpretation", the fact that (1.18) does not hold seems absurd. For is seems that any consequence relation should be *reflexive*: according to any consequence relation, inferring σ from σ should be correct. As $\vDash^{ts}$ is not reflexive, it does not qualify as a well-behaved consequence relation.

The *argument form* $\sigma \vDash \sigma$ is *classically* valid, i.e., valid in classical logic. With Liar sentences around, classical logic has, in some way or the other, to be modified. It seems a desirable property of a fixed point consequence relation that this modification—relative to classical logic—is rather small. Fact 1 has been put forward as an attractive selling point of STCT.

Fact 1 *Any argument form that is classically valid is* $\vDash^{st}$*-valid.*

Fact 1 distinguishes $\vDash^{st}$ from the other three fixed point consequence relations. We already saw that $\vDash^{ts}$ does not validate the argument form $\sigma \vDash \sigma$, which is classically valid. Although $\vDash^{ss}$ and $\vDash^{tt}$ validate argument form $\sigma \vDash \sigma$, they do not validate all classically valid argument forms. For instance, we have that:

$$\nvDash^{ss} \sigma \to \sigma \tag{1.19}$$

$$\alpha, \alpha \to \beta \nvDash^{tt} \beta \tag{1.20}$$

For illustrations of (1.19) and (1.20), the reader is referred to Section 7. Fact 1 certainly is a nice property for a consequence relation to have, and it may provide a reason for preferring $\vDash^{st}$ over $\vDash^{ss}$ and $\vDash^{tt}$. However, with Liar sentences around, any "good" property of a fixed point consequence relation comes at a price; the relation must also give up some intuitively plausible principles. The price that $\vDash^{st}$ has to pay for Fact 1 is:

Fact 2 $\vDash^{st}$ *is non-transitive:* $\alpha \vDash^{st} \beta$ *and* $\beta \vDash^{st} \gamma \not\Rightarrow \alpha \vDash^{st} \gamma$

In a nutshell, we may say that $\models^{st}$ saves classical logic (Fact 1) but has to give up its meta-theory (Fact 2). To see that $\models^{st}$ is non-transitive, consider a Liar sentence $\neg T(\lambda)$. As the Liar has value $\frac{1}{2}$ in any fixed point valuation, it is always tolerantly assertible but never strictly. As the Liar is always tolerantly assertible, we have that $\alpha \models^{st} \neg T(\lambda)$ for any sentence α whatsoever. So in particular, for the sentence 'snow is white'. As the Liar is never strictly assertible, we have that $\neg T(\lambda) \models^{st} \beta$ for any sentence β whatsoever. So, in particular, for the sentence 'snow is black'. Thus, according to $\models^{st}$, 'snow is white' implies the Liar, the Liar implies 'snow is black', but 'snow is white' does not imply 'snow is black'. Hence, $\models^{st}$ is not transitive.

In this thesis, we do not enter the discussion as to whether Fact 2 is too high a price to be paid for Fact 1. In Section 7, we are concerned with another, related, issue that has to be addressed by a STCT proponent. For $\models^{st}$ crucially relies on the distinction between strict and tolerant assertions and denials. As a consequence, STCT is committed to holding that there are *four* assertoric speech acts. As such, one may ask whether strict-tolerant distinction is, by itself, not all too costly. Ripley [46] has argued that it is not, as the strict-tolerant distinction is *not a primitive* one. In other words, the strict can be understood in terms of the tolerant, or, vice versa, the tolerant can be understood in terms of the strict. As we will see in Section 7, the combination of assertoric semantics and the strict-tolerant calculus provides a fruitful framework in which to assess this claim.

(22) From theories of truth to riddles about truth. We started our tour of the thesis with Sarah, who managed to pass the two guards safely by reasoning with the notion of truth. Let us, before we conclude our tour, return to Sarah and provide her with some guidance for further strolls through labyrinths. In the movie *Labyrinth*, Jareth (also known as the King of Goblins or David Bowie) tries to refrain Sarah from reaching his castle in the center of the labyrinth by confronting her with all kinds of riddles and puzzles. In the movie *Labyrinth: the sequel*, Jareth still has these bad habits. In particular, he sets up the *four roads riddle* (see also Section 3 and Section 4):

> There are four roads, numbered 1, 2, 3 and 4. Three out of these four roads lead to certain death, whereas the other road leads to the castle. The four roads are guarded by a single knight, and Sarah is only allowed to ask one yes-no question to the knight.

Confronted with the four roads riddle, Sarah is perplexed. For, how can she distinguish four possibilities by asking a single yes-no question? But then she remembers reading a story in which the four roads riddle was discussed. The moral of the story was formulated as a slogan: *self-referential truth has computational power* (see Section 4). She recalls from the story that asking the following yes-no question allows her to find out which road leads to the castle.

Is it the case that: (your answer to this very question is 'no' and the first door leads to the castle) or (your answer to this very question is 'yes' and the second door leads to the castle) or the third door leads to the castle?

Upon hearing Sarah's question, the knight needs some time to deliberate, but then he answers the question with 'I can *both* answer your question with 'yes' or

'no". From this answer, Sarah concludes that the second door leads to the castle. When confronted with Sarah's conclusion, the knight looks worried. Then, the following dialogue evolves.

Knight: How do you know that the second door leads to the castle?
Sarah: I remember reading a manuscript, called *Playing with Truth*. In Section 3 and Section 4 it was explained that when you answer with *both*, the second door leads to the castle.
Knight: I also read that manuscript. In the sections you are alluding to, it is assumed that I answer yes-no questions in accordance with the *strict* assertoric norm. However, the manuscript also considers a *tolerant* assertoric norm. If I answer in accordance with the tolerant norm, I may also answer your question with *both* when the first door leads to the castle. So then, what makes you think that I answered in accordance with the strict norm?
Sarah: (looks scared) I see... Indeed, I do not know which norms you try to live up to. However, it seems *plausible* that a knight, who always *speaks truly*, answers yes-no questions in accordance with the *strict* assertoric norm. Then again, as you're a knight, I can simply find out how knights behave by asking you a question about yourself. So then, do you answer in accordance with the strict norm?
Knight: ...
Jareth: Hahaha... He doesn't answer your questions anymore! With riddles about truth, it's all in the rules of the game: you were only allowed to ask a single question and you had your go. Bad luck with your decision!

As Sarah couldn't ask a second question, she decided to stick to her assumption that the knight had answered her question in accordance with the strict norm. Then, she acted in accordance with her assumption and passed through the second door. At that point, the movie ends.

(23) Playing with truth. We acted as follows. We started out with a *riddle about truth*, which we used to illustrate in which sense truth plays an expressive function. Then, we illustrated that this expressive function can be realized by a transparent notion of truth, i.e., a notion according to which 'x... is true' and x are, in all (non-opaque) contexts intersubstitutable for all declarative sentences x. We noted that the transparency of truth is a property of Strong Kleene (SK) theories of truth. Then, we introduced three frameworks in which to define and study theories of truth according to which truth is a transparent notion. A common feature of these frameworks is that truth is understood, in terms of the assertoric rules that govern it, as a primitive notion.

We introduced the notion of an English assertoric norm and showed how such norms can be formalized as closure conditions on branches of assertoric trees: i.e., we developed *assertoric semantics*. We considered two such closure conditions, which give rise to the theories of truth $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$. Both theories have a Strong Kleene character and satisfy the transparency of truth.

We asked whether the fact that $\mathcal{V}^t_M$ is compositional and $\mathcal{V}^s_M$ is not could be understood in terms of closure conditions. The *method of closure games* answers that question affirmatively: by putting constraints on the closure conditions of expansions ("fine-grained branches"), the method of closure games gives us a uniform approach to all (3- and 4-valued) SK fixed point valuations.

Assertoric semantics and the method of closure games are frameworks to define theories of truth. As such, the frameworks do not specify a consequence relation of the language for which the theories of truth are constructed. To study such consequence relations, we developed a signed tableau system, which we called the *strict-tolerant calculus*. To develop the calculus, we followed, o.a., Cobreros et al. [11], in drawing a distinction between *strict* assertions and denials and *tolerant* assertions and denials. Then, we exploited the distinction by equipping our calculus with four signs, which are naturally associated with the four different speech acts that arise out of the strict-tolerant distinction. We explained that the strict-tolerant calculus gives us a uniform approach to the SK fixed point consequence relations.

We introduced the notion of a *Generalized Strong Kleene fixed point valuation* and we pointed out that the method of closure games can be (indirectly) used to define such valuations. We explained how the notion of a Generalized Strong Kleene theory of truth suggests that a proposed desideratum for theories of truth has to be reconsidered.

Assertoric semantics and the strict-tolerant calculus are natural bedfellows. Assertoric semantics can be thought of as the semantic version of the strict-tolerant calculus. Dually, the strict-tolerant calculus can be thought of as the syntactic version of assertoric semantics. The connection between assertoric semantics and the strict tolerant calculus is nicely illustrated via $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$. These valuation functions of assertoric semantics can be induced by the closure conditions of the strict-tolerant calculus (modeling the *strict-tolerant norm*), augmented with the closure conditions associated with M. We explained that $\mathcal{V}^t_M$, $\mathcal{V}^s_M$ and the strict-tolerant calculus shed light on the *Strict Tolerant Conception of Truth*.

Then, we presented another riddle about truth, and we used this riddle to illustrate that there is a sense in which we can say that *self-referential truth has computational power*. Our last riddle gave rise to the same question that we asked after presenting our first riddle: *what makes a knight*? Sarah assumed that knights are made of strict assertoric norms.

In a nutshell, we acted as follows: we have been *Playing with Truth*.

1.2 Future work

Here are some possible directions of future work.

(A) Exploring the connections between $\mathcal{V}^t_M$, $\mathcal{V}^s_M$, the strict-tolerant calculus and STCT. As we illustrated in **(19)** and **(20)**, there are clear connections between the theories $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$ of assertoric semantics, the strict-tolerant calculus and STCT. More generally, STCT is a novel and, I take it, interesting and promising conception of truth, and the techniques developed in this thesis can be used to foster our understanding of it. For example, an interesting question to ask is whether our techniques can be used to define alternative consequence relations that satisfy the attractive highly classical behavior of the consequence relation that is advocated by STCT.

(B) Other frameworks for truth. This thesis is called 'Playing with Truth'. However, 'Playing with Strong Kleene Truth' also covers the subject of the the-

sis quite well. For, the *results* that are obtained with assertoric semantics, the method of closure games and the strict-tolerant calculus all have a Strong Kleene character. However, the main *techniques* of the three frameworks (put closure conditions on branches of assertoric trees, on expansions of closure games or on paths of tableaux) do, as such, not have this character. The question arises whether the techniques can also be used to deliver results that have a, say, *Weak Kleene* or *Supervaluation* character. Further, the question arises how our techniques and their motivation relates to other (non-Kripkean) frameworks for truth, such as, e.g., the Gupta-Belnap *revision theories of truth* (cf. [24]) or Gaifman's *Pointer Semanics* (cf. [18]).

(C) Expressive completeness. Kripke [33] criticized his own interpretation of the minimal fixed point along the following lines. With $\neg T(\lambda)$ the Liar, we have that $\mathcal{K}_M^{min}(\neg T(\lambda)) = \mathbf{n}$. As $\mathcal{K}_M^{min}(\neg T(\lambda)) \neq \mathbf{a}$, the Liar is not true. Yet when we express, in L_T, that the Liar is not true, i.e., by uttering $\neg T([\neg T(\lambda)])$, we are left with a sentence that we can't assert, as $\mathcal{K}_M^{min}(\neg T([\neg T(\lambda)])) = \mathbf{n}$. In one sentence, we may say that Kripke complained that $\mathcal{K}_M^{min}$ is *not expressive complete*. Very roughly, a theory of truth is expressive complete if it allows you to assert, in the object language, claims about the semantic values that are exploited by the meta-language. In this thesis, we were not concerned with the notion of expressive completeness at all. It seems interesting to study the notion of expressive completeness from the perspective of the assertoric conception of truth that is developed in this thesis. From this perspective, Kripke's criticism of $\mathcal{K}_M^{min}$ may have to be reconsidered. For, why would the fact that the Liar is not (strictly) assertible motivate you to claim that 'the Liar is not true' should be (strictly) assertible. On an assertoric conception of truth, strictly asserting 'the Liar is not true' is tantamount to strictly denying the Liar, and we know we can't do that. More generally, what do we mean with the expressive completeness of a theory of truth in an assertoric framework?

From a more technical point of view, it is interesting to see whether assertoric semantics or the method of closure games can be used to introduce semantic predicates within our language. For instance, can we specify appropriate rules and closure conditions for an *ungroundedness predicate*?

(D) Truth and arithmetic. In this thesis, sentential reference (i.e. reference to sentences) is modeled via *quotational languages* and the associated notion of a *ground model*. This approach is not uncommon, and is also found in the work of, a.o., Michael Kremer, Philip Kremer or Anil Gupta. However, another way of modeling sentential reference is by letting the truth language L_T to be an extension of the *language of arithmetic* L_A. On this approach, one first sets up a 1:1 correspondence G (a *Gödel numbering*) between the sentences of L_T and the natural numbers. As L_T extends L_A, it has available a *numeral* $\overline{n}$ which denotes, in the intended model $\mathbb{N}$ of L_A, the number n. Hence, relative to a Gödel numbering G and relative to $\mathbb{N}$, each sentence has a name in L_T: when $G(\sigma) = n$, the name for σ is $\overline{n}$.

However, the attractiveness of the arithmetic approach of sentential reference is not to be found on the semantic side, but rather on the syntactic side. For a sufficiently strong theory of arithmetic (such as *Robinson's arithmetic* or *Peano arithmetic*) *represents all recursive functions*. As the syntactic operations on sentences (formulas) of L_T are recursive functions, a sufficiently strong theory

of arithmetic can *represent the syntax of* L_T. In fact, a theory of arithmetic can also be thought of as a theory of the syntax of L_T. As an example, let RA be the sentences of Robinson's arithmetic and let G be a Gödel numbering. Clearly, to determine whether a sentence is a conjunction or not is a recursive operation. In other words, $\mathbf{CON} = \{n \mid G(n) \text{ is a conjunction}\}$ is a recursive set and hence, it is represented by RA, meaning that there will be a complex predicate of L_A, abbreviate it as Con, such that:

$$\mathsf{RA} \vdash \mathsf{Con}(\overline{n}) \Leftrightarrow n \in \mathbf{CON}$$

So, whenever n is (the Gödel number of) a conjunction, RA *proves* that it is a conjunction. In other words, RA represents the syntactic fact that n is conjunction. In a similar vein, RA represents all syntactic facts of L_T. This representation allows us to express (and evaluate) certain "laws of truth" in a language of arithmetic, such as the claim that a conjunction is true just in case its conjuncts are. On the arithmetical approach, self-reference is obtained via the *diagonalization theorem* of the arithmetical theory that is used to represent the syntax of L_T. In case of RA, the diagonalization theorem tells us that for any open formula $\phi(x)$ of L_T, there is a sentence σ of L_T such that $\mathsf{RA} \vdash \sigma \leftrightarrow \phi(\overline{\sigma})$, where $\overline{\sigma}$ is the numeral of the Gödel number of σ. Applying the theorem to $\neg T(x)$, it follows that there has to be some L_T sentence σ_λ such that:

$$\mathsf{RA} \vdash \sigma_\lambda \leftrightarrow \neg T(\overline{\sigma_\lambda})] \tag{1.21}$$

The sentence σ_λ represents the Liar on the arithmetical approach to sentential reference.

As the quotational language approach to self-reference as such does not allow us to represent the syntax of L_T, we may say that the arithmetical approach has a comparative advantage here. On the other hand, there is a sense in which the quotational language approach is more flexible. For instance, we can (as we do in Section 6) study ground models with and without *vicious reference* and explore the behavior of various theories of truth in such models. In contrast, on the arithmetical approach, there is always vicious reference around, as testified by (1.21).

Are there decisive reasons that force us to prefer, in general, one the two approaches to sentential reference over the other? Or do such reasons depend on the phenomenon under consideration? Besides these foundational questions, it is, at any rate, interesting to explore the techniques of this thesis in combination with an arithmetical approach to sentential reference. We hope to do so in future work.

(E) Game semantics The method of closure games is a method to equip L_T with a semantic valuation for L_T that is, ultimately, grounded in *game-theoretic concepts*. In particular, whether or not a sentence is assertible is determined by the existence of a *winning strategy* for one of the two players. As such, the method of closure games qualifies as a *game semantics*. It would be interesting to connect the method of closure games to other game semantics, such as the dialogue logics of Lorenzen and Lorenz [36] or the game semantics as developed by Hintikka [26]. More generally, our assertoric practice is often compared with a game. This suggest that the assertoric conception of truth as developed in this thesis can be backed up by a rationale that is similar to that underpinning

(versions of) game semantics.

(F) What makes a knight? As explained in **(22)**, Section 4 shows that self-referential truth has computational power. To obtain that result, we basically assumed that a knight answers a yes-no question σ? with 'yes', 'no', 'both' or 'neither' just in case $\mathcal{V}^s_M$ valuates σ as, respectively, **a**, **d**, **b** or **n**. Hence, we assumed that knights answer yes-no questions in accordance with the strict norm. There are various ways to alter the assumptions of Section 4. First I list some changes that, I think, are particulary interesting. Second, I explain why I think that the proposed changes are interesting in light of the results that are obtained in Section 4.

We may alter the form of a yes-no question that may be asked to a knight as follows. With $X \in \{A, D\}$ and $i \in \{s, t\}$, let the form of a yes-no question be equal to X^i_σ. As an example, A^t_σ corresponds to the question: do you tolerantly assert σ? The change in our questions naturally corresponds to a change in the manner in which a knight answers our questions. Previously, the knight answered the Liar question with 'neither', as $\mathcal{V}^s_M(\neg T(\lambda)) = \mathbf{n}$. This seems reasonable if we think of the Liar question as analogous to the question 'is your answer to this very question 'no'?'. However, when—in the to be developed framework under consideration—we ask a knight, say $A^s_{\neg T(\lambda)}$, we ask him whether he strictly asserts the Liar. Being a knight, he should answer with 'no'. Similarly, when we ask a knight $A^t_{\neg T(\lambda)}$, i.e., when we ask him whether he tolerantly asserts the Liar, he should answer with 'yes'. A further way in which we will alter the knight's answering function is by assuming that he always answers with either 'yes' or 'no' but that he does so probabilistically. For instance, with $T(\tau)$ the Truthteller, suppose that we ask the knight $A^s_{T(\tau)}$. As the knight is entitled, but not obliged, to strictly assert $T(\tau)$, he will answer 'yes' *with a certain probability*. But with which probability, i.e., for which x do we have that $p(A^s_{T(\tau)}, \text{yes}) = x$. One may say that, upon being asked $A^s_{T(\tau)}$ the knight takes a *random* decision as to whether or not to strictly assert the Truthteller. Hence, we can model the decision of a knight as a random variable over the courses of action that he may take. But then, should $p(A^s_{T(\tau)}, \text{yes}) = \frac{1}{3}$ as he may either strictly assert $T(\tau)$, tolerantly assert and tolerantly deny $T(\tau)$ or strictly deny $T(\tau)$? Or should $p(A^s_{T(\tau)}, \text{yes}) = \frac{1}{2}$, as, whenever possible, a knight should assert and deny strictly? Or does a tolerant assertion of the Truthteller not commit one to tolerantly deny it as well, in which case it seems plausible that $p(A^s_{T(\tau)}, \text{yes}) = \frac{1}{4}$. At any rate, it seems interesting to explore the probability spaces that emerge from different assumptions on the knight's behavior in the sketched "strict-tolerant framework".

In Section 4, we showed that the four roads riddle (cf. **(22)**) could be solved by asking a single self-referential yes-no question. We presented this result in computational terms. One way to object to this presentation is by responding that "on no natural notion of computational complexity, the four roads riddle is solvable in one question. So much the worse for the notion of computational complexity that is underlying your result". In Section 4, I respond to that objection by pointing out that the four roads riddle can also be solved in one question according to the notion of *quantum query complexity*, a notion of computational complexity that is used by researchers in *quantum computation*. Although the point of the "quantum reply" in Section 4 is merely

to rebut the considered objection, the question arises as to whether there are connections between the framework of Section 4 and quantum computation. An important distinction between classical and quantum computation is found in the *non-classical probability distributions* that are associated with the latter paradigm. It would be interesting if the probability spaces that are obtained in the sketched "strict-tolerant framework" connect, somehow, to the probability spaces that are used in quantum computation. Can quantum-like probability spaces arise out of self-referential truth? To explore this speculative question is the topic of future research.

1.3 Overview of the Thesis (Summary)

In the remainder of this thesis, we present the following papers:

Section 2: A Framework for Riddles about Truth that do not involve Self-Reference [58].

In this paper, we present a framework in which we analyze three riddles about truth that are all (originally) due to Smullyan. We start with *the riddle of the yes-no brothers* and then the somewhat more complicated *riddle of the da-ja brothers* is studied. Finally, we study the *Hardest Logic Puzzle Ever* ($HLPE$). We present the respective riddles as sets of sentences of *quotational languages*, which are interpreted by *sentence-structures*. Using a *revision-process* the consistency of these sets is established. In our formal framework we observe some interesting dissimilarities between $HLPE$'s available solutions that were hidden due to their previous formulation in natural language. Finally, we discuss more recent solutions to $HLPE$ which, by means of *self-referential questions*, reduce the number of questions that have to be asked in order to solve $HLPE$. Although the essence of the paper is to introduce a framework that allows us to formalize riddles about truth that do not involve self-reference, we will also shed some formal light on the self-referential solutions to $HLPE$.

Section 3: On the Behavior of True and False [65].

Uzquiano showed that the Hardest Logic Puzzle Ever ($HLPE$) (in its amended form due to Rabern and Rabern) has a solution in only two questions. Uzquiano concludes his paper by noting that his solution strategy naturally suggests a harder variation of the puzzle which, as he remarks, he does not know how to solve in two questions. Wheeler and Barahona formulated a three question solution to Uzquiano's puzzle and gave an information theoretic argument to establish that a two question solution for Uzquiano's puzzle does not exist. However, their argument crucially relies on a certain conception of what it means to answer *self-referential* yes-no questions *truly* and *falsely*. We propose an alternative such conception which, as we show, allows one to solve Uzquiano's puzzle in two questions. The solution strategy adopted suggests an even harder variation of Uzquiano's puzzle which, as we will show, can also be solved in two questions. Just as all previous solutions to versions of $HLPE$, our solution is presented informally. The second part of the paper investigates the prospects of formally representing solutions to $HLPE$ by exploiting theories of truth.

Section 4: Assertoric Semantics and the Computational Power of Self-

Referential Truth [57].

There is no consensus as to whether a Liar sentence is meaningful or not. Still, a widespread conviction with respect to Liar sentences (and other *ungrounded* sentences) is that, whether or not they are meaningful, they are *useless*. The philosophical contribution of this paper is to put this conviction into question. Using the framework of *assertoric semantics*, which is a semantic valuation method for languages of self-referential truth that has been developed by the author, we show that certain computational problems, called *query structures*, can be solved more efficiently by an agent who has self-referential resources (amongst which are Liar sentences) than by an agent who has only classical resources; we establish the *computational power of self-referential truth*. The paper concludes with some thoughts on the implications of the established result for deflationary accounts of truth.

Section 5: From Closure Games to Generalized Strong Kleene Theories of Truth [63].

In this paper, we study *the method of closure games*, which is a game theoretic valuation method for languages of self-referential truth, developed by the author. We prove two theorems which jointly establish that the method of closure games characterizes all 3- and 4-valued Strong Kleene theories of truth (SK theories) in a uniform manner. Another theorem states conditions under which SK theories can be combined into *Generalized Strong Kleene theories of truth* (GSK theories). In contrast to a SK theory, a GSK theory recognizes more than one sense of strong assertibility—where a sentence is strongly assertible just in case it is assertible and its negation is not. Exploiting the relations between SK theories laid bare by the method of closure games, we then show how to define 5-, 6-, 7-, 8- and 10-valued GSK theories.

Section 6: Alternative Ways for Truth to Behave when there's no Vicious Reference [60].

In a recent paper, Philip Kremer proposes a formal and theory-relative desideratum for theories of truth that is spelled out in terms of the notion of 'no vicious reference'. Kremer's *Modified Gupta-Belnap Desideratum* (**MGBD**) reads as follows: if theory of truth **T** dictates that there is no vicious reference in ground model M, then **T** should dictate that truth behaves like a classical concept in M. In this paper, we suggest an alternative desideratum (**AD**): if theory of truth **T** dictates that there is no vicious reference in ground model M, then **T** should dictate that all T-sentences are (strongly) assertible in M. We illustrate that **MGBD** and **AD** are not equivalent by means of a *Generalized Strong Kleene theory of truth* and we argue that **AD** is preferable over **MGBD** as a desideratum for theories of truth.

Section 7: Strict-Tolerant Tableaux for Strong Kleene Truth [66].

We discuss four distinct semantic consequence relations which are based on Strong Kleene theories of truth and which generalize the notion of classical consequence to 3-valued logics. Then we set up a uniform signed tableau calculus (the *strict-tolerant calculus*) which we show to be sound and complete with respect to each of the four semantic consequence relations. The signs employed by our calculus are A^s, D^s, A^t and D^t which indicate a *strict assertion*, *strict denial*, *tolerant assertion* and *tolerant denial* respectively. Recently, Ripley ap-

plied the strict-tolerant account of assertion and denial (originally developed by Cobreros et all. to bear on vagueness) to develop a new approach to truth and alethic paradox, which we call the *Strict Tolerant Conception of Truth* (STCT). The paper aims to contribute to our understanding of STCT in at least three ways. First, by developing the strict-tolerant calculus. Second, by developing a semantic version of the strict-tolerant calculus (*assertoric semantics*) which informs us about the (strict-tolerant) assertoric possibilities relative to a fixed *ground model*. Third, by showing that the strict-tolerant calculus and assertoric semantics jointly suggest that STCT's claim that "the strict and tolerant can be understood in terms of one another" has to be reconsidered. The paper concludes with a methodological comparison between the strict-tolerant calculus and other calculi that are also sound and complete with respect to (some of the) semantic consequence relations based on Strong Kleene theories of truth.

Section 8: A Calculus for Belnap's Logic in Which Each Proof Consists of Two Trees[19] [67].

In this paper we introduce a Gentzen calculus for (a functionally complete variant of) Belnap's logic in which establishing the provability of a sequent in general requires *two* proof trees, one establishing that whenever all premises are true some conclusion is true and one that guarantees the falsity of at least one premise if all conclusions are false. The calculus can also be put to use in proving that one statement *necessarily approximates* another, where necessary approximation is a natural dual of entailment. The calculus, and its tableau variant, not only capture the classical connectives, but also the 'information' connectives of four-valued Belnap logics. This answers a question by Avron.

[19] This paper is joint work with Reinhard Muskens.

Chapter 2

A Framework for Riddles about Truth that do not involve Self-Reference

2.1 Abstract

In this paper, we present a framework in which we analyze three riddles about truth that are all (originally) due to Smullyan. We start with *the riddle of the yes-no brothers* and then the somewhat more complicated *riddle of the da-ja brothers* is studied. Finally, we study the *Hardest Logic Puzzle Ever* ($HLPE$). We present the respective riddles as sets of sentences of *quotational languages*, which are interpreted by *sentence-structures*. Using a *revision-process* the consistency of these sets is established. In our formal framework we observe some interesting dissimilarities between $HLPE$'s available solutions that were hidden due to their previous formulation in natural language. Finally, we discuss more recent solutions to $HLPE$ which, by means of *self-referential questions*, reduce the number of questions that have to be asked in order to solve $HLPE$. Although the essence of the paper is to introduce a framework that allows us to formalize riddles about truth that do not involve self-reference, we will also shed some formal light on the self-referential solutions to $HLPE$.

2.2 Introduction

In this paper, I analyze a cluster of riddles about truth that are all (originally) due[1] to Smullyan ([49]). Although most of Smullyan's riddles are trapped in the fun-logic cage, one of them was set free by George Boolos ([9]) and now lives in academia. Upon setting it free in 1996, Boolos baptized the riddle 'the Hardest Logic Puzzle Ever' ($HLPE$). For the readers not familiar with $HLPE$, here is its formulation due to Boolos:

[1] According to Raymond Smullyan however, the 'Hardest Logic Puzzle Ever', though attributed to him by George Boolos, is not due to him. Professor Smullyan doesn't know to whom the puzzle is due.

> The Puzzle: Three gods A, B and C are called, in some order, True, False, and Random. True always speaks truly, False always speaks falsely, but whether Random speaks truly or falsely is a completely *random* matter. Your task is to determine the identities of A, B, and C by asking three yes-no questions; each question must be put to exactly one god. The gods understand English, but will answer all questions in their own language, in which the words for 'yes' and 'no' are 'da' and 'ja' in some order. *You do not know which word means which.* Before I present the somewhat lengthy solution, let me give answers to certain questions about the puzzle that occasionally arise:
>
> - (B1) It could be that some god gets asked more than one question (and hence that some god is not asked any question at all)
> - (B2) What the second question is, and to which god it is put, may depend on the answer to the first question. (And of course similarly for the third question)
> - (B3)Whether Random speaks truly or not should be thought of as depending on the flip of a coin hidden in his brain: if the coin comes down heads, he speaks truly, if tails, falsely.
> - (B4)Random will always answer 'da' or 'ja' when asked any yes-no question. (Boolos [9, p62])

Until very recently, $HLPE$ has lived a quiet life in academia. We only find Tim Roberts ([47]) coming up with an alternative solution for $HLPE$, criticizing Boolos' solution as being 'unnecessarily complicated'. Recently, Rabern and Rabern ([43]) interestingly observed an ambiguity in Boolos' instruction for $HLPE$ concerning the behavior of Random. They observed that there is a genuine difference between the following two courses of action that may be followed by Random upon being addressed some question **Q**.

1. Random flips a coin and then, depending on the outcome of the coin-flip, answers **Q** with either 'da' or 'ja'.

2. Random flips a coin and then, depending on the outcome of the coin-flip, answers **Q** either truly or falsely.

We refer to the first course of action as the *syntactic protocol* and to the second course of action as the *semantic protocol*. Rabern and Rabern point out that 'most commentators on the puzzle [i.e. $HLPE$] have assumed that Random answers randomly and that therefore nothing can be gleaned from his answers; but that is not how Random works'. They claim that previous commentators have assumed that Random's answers are completely unpredictable and they show that this assumption is false when Random's behavior is understood in line with Boolos' instruction ($B3$), that is, according to the semantic protocol. The syntactic protocol is introduced to make explicit the behavior of Random that, arguably, Boolos intended to describe. This leaves us with two different versions of $HLPE$, one in which Random behaves according to the syntactic protocol, denoted $HLPE_{syn}$, and one in which Random behaves according to the semantic protocol, denoted $HLPE_{sem}$.

Rabern and Rabern exploit their observations concerning the behavior of Random to give a (three-question) solution to $HLPE_{sem}$ that differs substantially from the previous solutions due to Boolos and Roberts. Interestingly, they also show that $HLPE_{sem}$ can be solved by asking only two questions (!) when we are allowed to ask the gods *self-referential questions*. As observed by Uzquiano ([53]), the results of Rabern and Rabern prompt the question whether $HLPE_{syn}$, which is, after all, more properly called "the hardest logic puzzle ever", allows for a two-question solution as well. To answer this question, Uzquiano observes that a refinement of the puzzle is in place. For, the omniscience of True and False is somewhat at odds with the random behavior of Random; can an omniscient god predict a random event? Two answers suggest themselves:

1. True and False have the ability to predict the outcome of the coin flip in Random's brain (as they are omniscient).

2. True and False do not have the ability to predict the outcome of the coin flip in Random's brain (as the coin flip is a random event).

As indicated by the formulation of the two answers, the (in)ability of True and False to predict the outcome of random coin flips is independent of the distinction between semantic and syntactic $HLPE$. Thus, we obtain four versions of $HLPE$; $HLPE_{syn}$ splits into $HLPE^{omn}_{syn}$, in which True and False can predict random events, and $HLPE^{ran}_{syn}$, in which they do not have this ability. Likewise, $HLPE_{sem}$ splits into $HLPE^{omn}_{sem}$ and $HLPE^{ran}_{sem}$. Uzquiano shows that both $HLPE^{omn}_{syn}$ and $HLPE^{ran}_{syn}$ allow for a two question solution. However, the nature of his two solutions is fundamentally distinct, as his solution to $HLPE^{omn}_{syn}$ depends on self-referential questions, while his solution to $HLPE^{ran}_{syn}$ does not. The distinction between $HLPE^{omn}_{sem}$ and $HLPE^{ran}_{sem}$ is less important; the solution of Boolos, of Roberts and the solutions of Rabern and Rabern work for both versions of $HLPE_{sem}$. For that reason, we will not be concerned with $HLPE^{ran}_{sem}$, but only with the other three versions of $HLPE$.

A common feature of all mentioned solutions for $HLPE$, in[2] one of its variants, is their being fully couched in natural language; there is no formalization of the reasoning involved that leads to the respective solution for $HLPE$. *The central theme in this paper is the formalization of such reasoning.* The paper is organized as follows.

Section 2.3 We provide our formal framework, consisting of the use of *quotational languages* and *sentence-structures*. Before studying $HLPE$ (which we do in Section 2.4), we confront our framework with two other riddles due to Smullyan, whose essential features are also present in $HLPE$; the *riddle of the yes-no brothers* and the *riddle of the da-ja brothers*. In analyzing them, we present sets of first order axioms that represent the riddles and show how their solutions can be derived from these axioms. The presented sets of axioms contain an axiom scheme that closely resembles the infamous T-scheme. For that reason, the consistency of our axioms is under suspicion. However, borrowing from Gupta ([23]) the consistency of the axioms can be established by explicitly

[2]From now on '$HLPE$', without sub or superscripts, will be used as shorthand for 'one of the variants of the hardest logic puzzle ever'.

constructing a model for them using a *revision process*. For each of the two riddles, we present two different solutions; one solution having *degree* 0 and the other having degree 1, where the degree of a sentence is the number of embedded quotations occurring in that sentence. The degree 0 solution for the riddle of the yes-no brothers may be obtained as an application of Smullyan's *fundamental principle* ([49]). A key feature of Smullyan's puzzles is that the correctness of a solution is easily verified while it is much harder to come up with a solution. For that reason, we present the *possible worlds method*, allowing us to obtain the degree 0 solution (and in fact Smulllyan's fundamental principle) in a constructive manner. We also use the possible worlds method to obtain the degree 0 solution for the riddle of the da-ja brothers.

Section 2.4 Then it is time to study the solutions to $HLPE$ which do not rely on the possibility to ask self-referential questions. In Section 2.4.1 we present the natural language solutions due to Boolos and Roberts and we show that they have the same three stage structure. In Section 2.4.2, we formalize $HLPE^{omn}_{syn}$ and show how the solutions of Boolos (having degree 0) and Roberts (having degree 1) can be derived from our representation of $HLPE^{omn}_{syn}$. In Section 2.4.3, we will discuss the three-question solution offered by Rabern and Rabern for $HLPE^{omn}_{sem}$, i.e., for the semantic version of $HLPE$ where True and False can predict random events. Section 2.4.4 is concerned with (the formalization) of $HLPE^{ran}_{syn}$. Although the solutions of Boolos' and Roberts' for $HLPE^{omn}_{syn}$ carry over to $HLPE^{ran}_{syn}$, we reveal an interesting dissimilarity in their solutions: for Roberts' solution to go through, it is necessary that the gods know their own answers to the questions, whereas this is not a necessary condition for Boolos' solution. In Section 3.5, we discuss Uzquiano's two-question solution for $HLPE^{ran}_{syn}$, and show that a formal derivation of this solution demands, when compared Boolos' and Roberts' solutions for $HLPE^{ran}_{syn}$, some additional axioms.

Section 2.5 Here, we discuss the self-referential solutions to $HLPE$ and we sketch a possible way to formalize these solutions as well.

2.3 Two riddles due to Smullyan

2.3.1 The riddle of the yes-no brothers

In this section, I present a logic riddle that is originally due to Raymond Smullyan ([49]). However, the version of the riddle that I present is taken from the movie 'the Labyrinth' ([25]). Let me state the riddle.

There are two brothers, b_L and b_T, having the following remarkable properties. One of them is always lying (b_L), the other is always speaking the truth (b_T). The brothers know this of each other. That is, b_T knows that b_L is always lying and b_L knows that b_T is always speaking the truth. 'Speaking' is probably not the right word to characterize the linguistic behavior of the brothers, in fact, they hardly speak at all. Their only linguistic behavior consists of giving yes / no answers to questions in sentential form. Thus, if we ask 'snow is white?' to b_T, he will answer 'yes', whereas b_L's answer to this question will be 'no'. A last remarkable feature of the brothers is as follows. When a single question is asked to and answered by either one of them, they both cease to exist. You are traveling along a road and suddenly the road forks. You have to continue your journey by either heading left (l) or right (r). One of the roads, call

it 'the good road', leads you to the destination of your travel. However, taking the other road (the bad one) will result in a vicious death. Thus, taking the good road is of crucial importance. Unfortunately, you have no clue which road is the good one. But at the cross-roads, the two brothers are stationed, and each of them knows which road is the good road. You know this and you also know the brothers' remarkable properties, but you do not know which brother is the liar and which brother is the truth-speaker. The riddle of the yes-no brothers is as follows.

Given the circumstances just sketched, can you reach the destination of your travel with certainty?

When the brothers lack the 'jointly ceasing to exist after being addressed a single question' property, the riddle can hardly be called a 'riddle' anymore. You just pick a brother, take a sentence σ of which you know that it is true and of which you know that the brothers know that it is true and ask 'σ'?. By this, you find out whether you are dealing with b_L or b_T. Next, you ask 'l is the good road?' to (say) b_T and depending on whether his answer is 'yes' or 'no', you continue your journey by taking road l or r respectively. This strategy is unavailable, for the brothers jointly cease to exist after one of them has answered a question. Nonetheless, as you may have already observed, the riddle can be solved by addressing $\mathbf{Q_1}$ to either one of the brothers:

$\mathbf{Q_1}$*: The answer of your brother to the question 'l is the good road?' is 'yes' ?*

We leave the verification that $\mathbf{Q_1}$ solves the riddle to the reader. It turns out that, although the brothers cease to exist after one of them has answered a question, we can safely continue our journey by asking $\mathbf{Q_1}$ to either one of the brothers. When the answer we get on $\mathbf{Q_1}$ is 'no', we take road l, and when the answer we get is 'yes' we take road r.

2.3.2 Modeling the riddle of the yes-no brothers

In the previous section, our reasoning led us to the following conclusion:

Conclusion: 'The left road is good if and only if the answer to $\mathbf{Q_1}$ of either one of the brothers is 'no', and the right road is good if and only if the answer to $\mathbf{Q_1}$ of either one of the brothers is 'yes' '.

An adequate treatment of the riddle should formalize the reasoning that leads us to this conclusion. That is, we want to have a set of axioms, describing the riddle, and show how we can infer the above conclusion from these axioms. As the informal reasoning in Section 2.3.1 employed the concept of knowledge (e.g. 'b_T *knows* that ...') one might think of introducing a modal language —interpreting the modal operators as epistemic operators — to fulfill our task. However, in this paper we only work with first order languages. In fact, the riddle of the yes-no brothers is modeled in a first order language that does not even poses a knowledge predicate. As we shall see, an adequate treatment of $HLPE^{ran}_{syn}$ (Section 2.4.4) demands that our object-language does contain a knowledge predicate.

In order to describe the axioms for the riddle of the yes-no brothers, we shall introduce, in Section 2.3.4, the first order language $\mathcal{L}_B^{[\cdot]}$. Before doing so, we first have a section containing some technical preliminaries.

2.3.3 The formal framework: quotational languages and sentence-structures.

I use 'first order language' as shorthand for 'first order language with identity'. That is, '=' is taken to be a *logical symbol*, denoting the identity relation. That being said, first order languages will be identified with their non-logical vocabulary. When $\mathcal{L}$ is a first order language, we use $Sen(\mathcal{L})$ for the set of all sentences of $\mathcal{L}$. Thus, we assume that the reader is familiar with the standard syntactic operations that build sentences out of the symbols of a first order language.

In our model of the riddle, we make use of a *quotational* language. The reason that we work with quotational languages is that doing so allows us to refer to sentences in our object language. For instance, with respect to the riddle of the yes-no brothers, it allows us to formalize expressions like: 'the lying brother answers 'no' to 'snow is white''. A natural way to refer to sentences is by quoting them, as we just did, and a quotational language allows us to capture this type of sentential reference. Formally, a language is quotational if it is obtained from some base-language by means of the process of *quotational closure*. The process is defined as follows:

Definition 2.1 Quotational language, quotational closure.
Let $\mathcal{L}$ be an arbitrary first order language. We set $\mathcal{L}^0 = \mathcal{L}$ and define:

- $\mathcal{L}^{n+1} = \mathcal{L}^n \cup \{[\sigma] \mid \sigma \in Sen(\mathcal{L}^n)\}$, $n \geq 0$
- $\mathcal{L}^{[\cdot]} = \bigcup_{i \geq 0} \mathcal{L}^i$

When σ is a sentence of $\mathcal{L}^n$, $[\sigma]$ is a *constant symbol* of $\mathcal{L}^{n+1}$. We say that $\mathcal{L}^{[\cdot]}$ is the *quotational language* obtained from $\mathcal{L}$ by means of (the process of) *quotational closure*. The hierarchy $\mathcal{L}^0, \mathcal{L}^1, \ldots$ is called the *quotational hierarchy*. □

Thus, a quotational language $\mathcal{L}^{[\cdot]}$ has a *canonical* constant symbol $[\sigma]$ for each sentence σ. Canonical, in the sense that in any *sentence structure* $M = \langle D, I \rangle$ for $\mathcal{L}^{[\cdot]}$, we have that $I([\sigma]) = \sigma$. The notion of a sentence-structure is defined as follows.

Definition 2.2 Sentence-structures.
Let $\mathcal{L}$ be an arbitrary first order language and let $\mathcal{L}^{[\cdot]}$ be the quotational language obtained from $\mathcal{L}$ by quotational closure. A *sentence-structure* $M = \langle D, I \rangle$ for $\mathcal{L}^{[\cdot]}$ is a structure for $\mathcal{L}^{[\cdot]}$ such that:

1. $Sen(\mathcal{L}^{[\cdot]}) \subseteq D$.
2. $I(c) \notin Sen(\mathcal{L}^{[\cdot]})$ for any constant symbol $c \in \mathcal{L}$.
3. $I(f)(d_1, \ldots, d_n) \notin Sen(\mathcal{L}^{[\cdot]})$ for any n-place function symbol f and any sequence $d_1, \ldots d_n \in D$.

4. $I([\sigma]) = \sigma \in Sen(\mathcal{L}^{[\cdot]})$ for each constant symbol $[\sigma] \in \mathcal{L}^{[\cdot]}$. □

So in a sentence-structure for a quotational language $\mathcal{L}^{[\cdot]}$ there is for each sentence σ of $\mathcal{L}^{[\cdot]}$ *exactly one term* in the language $\mathcal{L}^{[\cdot]}$ (the canonical constant symbol $[\sigma]$) that refers to σ.

Note that each sentence σ of $\mathcal{L}^{[\cdot]}$ has a *first occurrence* in the hierarchy $Sen(\mathcal{L}^0)$, $Sen(\mathcal{L}^1)$, ... The level of this first occurrence is called the *degree* of σ.

Definition 2.3 The degree of a sentence.
Let $\mathcal{L}$ be an arbitrary first order language and let $\mathcal{L}^{[\cdot]}$ be the quotational language obtained from $\mathcal{L}$ by quotational closure. Let $\sigma \in Sen(\mathcal{L}^{[\cdot]})$. The *degree* of σ equals $n(\geq 0)$ if and only if $\sigma \in Sen(\mathcal{L}^n)$ and $\sigma \notin Sen(\mathcal{L}^{n-1})$. □

2.3.4 The language $\mathcal{L}_B^{[\cdot]}$ and a formal solution for the riddle

We now define the language $\mathcal{L}_B^{[\cdot]}$, 'the quotational language of the Brothers' mentioned at the end of Section 2.3.2.

Definition 2.4 The language $\mathcal{L}_B^{[\cdot]}$.
Let $\mathcal{L}_B = \{l, r, c_n, c_y, b_T, b_L, b_1, b_2, f_A, G\}$, where the first 8 symbols are constant symbols, where f_A is a binary function symbol and where G is a unary predicate symbol. $\mathcal{L}_B^{[\cdot]}$ is the language obtained by quotational closure from $\mathcal{L}_B$. □

The *intended interpretation* of $\mathcal{L}_B^{[\cdot]}$ can be 'read off from its symbolism' so to speak. In the intended interpretation, l refers to the left road, r refers to the right road, c_y refers to 'yes', c_n refers to 'no', b_T refers to the truth telling brother and b_L refers to the lying brother. The constant symbols b_1 and b_2 are alterative names, *in some order* for the lying and the truth telling brother. For instance, one may think of b_1 as denoting the brother "standing on the left", while b_2 refers to the brother "standing on the right". As we will see, the names b_1 and b_2 replace the indexical phrases 'you' and 'your brother' in the natural language solutions to the riddle. The predicate symbol G is interpreted as 'being the good road' and the intended interpretation of the binary function symbol f_A is as follows. Let c_1, c_2, c_3 be constant symbols of $\mathcal{L}_B^{[\cdot]}$. Then:

'$f_A(c_1, c_2) = c_3$' is interpreted[3] as 'the answer of c_1 to c_2 is c_3'.

Next, we use $\mathcal{L}_B^{[\cdot]}$ to describe the riddle via the theory $K \subseteq Sen(\mathcal{L}_B^{[\cdot]})$, which consists of the following axioms / axiom schemes. Below, and in the rest of the paper, all axiom schemes are understood to range over all the sentences of the quotational language under consideration.

Definition 2.5 The set of axioms K.
K is defined as the following theory in $\mathcal{L}_B^{[\cdot]}$:

$K_1 : G(l) \leftrightarrow \neg G(r)$.

[3]Thus, we have 'garbage sentences' as e.g. 'the answer of the left road to the right road is the truth-teller'. I do not regard it is a defect of our approach, as the syntax of natural language allows for similar constructions.

$K_2 : (b_1 = b_T \wedge b_2 = b_L) \vee (b_1 = b_L \wedge b_2 = b_T)$.
$K_3 : c_y \neq c_n$
$K_4 : f_A(b_T, [\sigma]) = c_y \vee f_A(b_T, [\sigma]) = c_n$
$K_5 : f_A(b_T, [\sigma]) = c_y \leftrightarrow f_A(b_L, [\sigma]) = c_n$
$K_6 : \sigma \leftrightarrow f_A(b_T, [\sigma]) = c_y$ □

The content of the axioms is clear from the intended interpretation of $\mathcal{L}_B^{[\cdot]}$. Note that the axioms tell us that the brothers are *omniscient*. That is, *every* true sentence of $\mathcal{L}_B^{[\cdot]}$ is answered correctly by the truth-teller and is answered falsely by the liar.

A solution to the riddle is a single question which allows us to find out which road is good. Assuming that the question is addressed to b_1, a solution to the riddle is a sentence σ of $\mathcal{L}_B^{[\cdot]}$ such that (2.1) and (2.2) can be inferred[4] from K:

$$G(r) \leftrightarrow f_A(b_1, [\sigma]) = c_y \tag{2.1}$$

$$G(l) \leftrightarrow f_A(b_1, [\sigma]) = c_n \tag{2.2}$$

A person who knows K and who is able to carry out inferences in first order logic can solve the riddle, as the following theorem shows.

Proposition 2.1 Solving the riddle.
Let K be as in Definition 2.5 and let $\Gamma \vdash_K \sigma$ mean that there is a derivation of σ from Γ using the inference rules of first order logic and the sentences in K as axioms. It holds that:

$\vdash_K G(r) \leftrightarrow f_A(b_1, [f_A(b_2, [G(l)]) = c_y]) = c_y$
$\vdash_K G(l) \leftrightarrow f_A(b_1, [f_A(b_2, [G(l)]) = c_y]) = c_n$

Proof: Left to the reader. □

We gave a set of axioms K, modeling the knowledge of a person who is invited to solve the riddle, and showed how such a person can solve the riddle; this is modeled by the derivation from K exemplified in Theorem 2.1. Are we done? Have we successfully modeled the semantic phenomenon under consideration? When K is an inconsistent set of axioms, our derivation comes down to a mere triviality, for then we can derive *any* sentence of $\mathcal{L}_B^{[\cdot]}$ from K. And when it comes to inconsistency, K is suspect. The reason for this is that K's axiom scheme K_6 closely resembles the well-known T-scheme. When $\mathcal{L}$ is an interpreted (first order) language, the T-scheme for $\mathcal{L}$ is as follows:

$$T\text{-}scheme : \sigma \leftrightarrow T(\langle\sigma\rangle) \text{ for all } \sigma \in \mathcal{L}$$

Here, T is a truth predicate and $\langle\sigma\rangle$ is any closed term of $\mathcal{L}$ that refers to σ (or a code for σ) in the intended interpretation. From the literature on truth, it is well known that a theory Δ formulated in a language $\mathcal{L}$ such that Δ represents the syntax of $\mathcal{L}$ and such that Δ derives the T-scheme cannot be consistent. Although K does not represent the syntax of $\mathcal{L}_B^{[\cdot]}$, it *does* derive a T-scheme in

[4] Of course, interchanging c_y and c_n in (2.1) and (2.2) also constitutes a solution. Further, (2.1) and (2.2) are equivalent with respect to K.

disguise. The last feature turns K into a suspect of inconsistency and unless we succeed in proving its innocence, our Theorem 1 is no satisfactory solution to the riddle. In the appendix, I prove that K is innocent. The proof that K is consistent is a modification of a proof by Gupta ([23]). Basically, the result obtained by Gupta was that quotational languages which are weak in their expressive resources, can consistently contain the T-scheme. In the appendix, we show how Gupta's proof can be adapted to construct a classical model for K, thus proving its consistency. The techniques employed there can also be used to prove the consistency of all other theories that will be considered in this paper.

2.3.5 Alternative solutions and the fundamental principle

In Section 2.3.4, we solved the riddle of the yes-no brothers by asking the question '$f_A(b_2, [G(l)]) = c_y$', to b_1, which is the formal analogue of question $\mathbf{Q_1}$ that was discussed in Section 2.3.1. Thus, in the terminology introduced by Definition 3, we solved the riddle by a sentence of degree 1. An interesting question to ask is for which degrees we can find solutions to the riddle. More precisely, we want to know for which n we can find $\sigma \in \mathcal{L}_B^n$ such that:

$$\vdash_K G(r) \leftrightarrow f_A(b_1, [\sigma]) = c_y \tag{2.3}$$

To construct sentences of arbitrary high degree ($n \geq 1$) that do the job is trivial. As the answer of the addressed brother is determined solely by the truth or falsity of the sentence asked, we can, as the reader may verify, 'upgrade' the sentence '$f_A(b_2, [G(l)]) = c_y$' by taking its conjunction with tautologies of arbitrary high degree. The remaining question is thus whether there exists a sentence of degree 0 that solves the riddle. There do exist such solutions. Solutions of degree 0 correspond with instances of what Smullyan ([49]) calls *the fundamental principle*. Let us quote the puzzle-master on the fundamental principle; remember that for Smullyan a knight is always speaking the truth whereas a knave always lies.

> The last two problems imply a very important principle well known to 'knight-knave' experts. As seen in the solutions of the last two problems, if P is any statement at all, whose truth or falsity you wish to ascertain, if a person known to be a knight or knave knows the answer to P, then you can find out from him in just one question whether P is true or false. You just ask him, 'Is the statement that you are a knight equivalent to the statement that P is true?' If he answers 'Yes' then you know that P is true; if he answers 'No', then you know that P is false. This principle will be used in the solution of the next three problems; we shall refer to it as the *fundamental principle*. (Smullyan [49, p126])

Applied to our riddle, Smullyan's fundamental principle teaches us that the following question $\mathbf{Q_0}$ allows us to reach the destination of our travel with certainty:

$\mathbf{Q_0}$: You are the truth-speaker if and only if left is the good road?

When we translate $\mathbf{Q_0}$ in $\mathcal{L}_B^{[\cdot]}$, we can show (as the reader may wish to verify) that we have a solution of degree 0:

$$\vdash_K G(r) \leftrightarrow f_A(b_1, [G(l) \leftrightarrow b_1 = b_T]) = c_n$$
$$\vdash_K G(l) \leftrightarrow f_A(b_1, [G(l) \leftrightarrow b_1 = b_T]) = c_y$$

We thus arrived at a solution of degree 0 by translating the solution $\mathbf{Q_0}$ (obtained from Smullyan's fundamental principle) and by realizing that the translation Q_0 is of degree 0. In fact, K derives Smullyan's fundamental principle.

Proposition 2.2 Fundamental 0-principle
Let $\sigma \in Sen(\mathcal{L}_B^{[\cdot]})$. Then:

1. $\vdash_K \sigma \leftrightarrow f_A(b_1, [b_1 = b_T \leftrightarrow \sigma]) = c_y$
2. $\vdash_K \neg\sigma \leftrightarrow f_A(u, [b_1 = b_T \leftrightarrow \sigma]) = c_n$

Proof : Left to the reader. □

We refer to Proposition 2.2 as the fundamental 0-principle because from it, the degree 0 solution to the riddle of the yes-no brothers immediately follows. $\mathbf{Q_1}$, the degree 1 solution to the riddle presented in Section 2.3.1, follows immediately from the fact that for every $\sigma \in Sen(\mathcal{L}_B^{[\cdot]})$ we have that:

$$\vdash_K \sigma \leftrightarrow f_A(b_1, [f_A(b_2, [\sigma]) = c_y]) = c_n$$
$$\vdash_K \neg\sigma \leftrightarrow f_A(b_1, [f_A(b_2, [\sigma]) = c_y]) = c_y$$

We can also obtain a degree 1 solution by asking a brother to reflect on his own habits. Let us call this result the fundamental 1-principle.

Proposition 2.3 Fundamental 1-principle
Let $\sigma \in Sen(\mathcal{L}_B^{[\cdot]})$. Then:

1. $\vdash_K \sigma \leftrightarrow f_A(b_1, [f_A(b_1, [\sigma]) = c_y]) = c_y$
2. $\vdash_K \neg\sigma \leftrightarrow f_A(b_1, [f_A(b_1, [\sigma]) = c_y]) = c_n$

Proof: Left to the reader. □

Proposition 2.3 is a formal analogue of what Rabern and Rabern ([43]) call the "Embedded Question Lemma". In natural language, Proposition 2.3 says that we can reveal the truth-value of σ by addressing the following question to a brother:

Is your answer to question 'σ' 'yes'?

2.3.6 The riddle of the da-ja brothers.

In this section, we introduce and discuss the riddle of the da-ja brothers. It is stated as follows.

Suppose that the two brothers of Section 2.3.1, while they understand English, answer all yes-no questions with the words 'da' and 'ja' which mean 'yes' and 'no', but not necessarily in that order. You do not know which word means which. Besides this

curiosity, the story is as in Section 2.3.1; you want to know whether the left or right road is good. Again, one of the brothers is a liar, the other is a truth speaker and you do not know which brother is which. Under these circumstances, can you find out which road is good by asking a single yes-no question?

To formalize the riddle of the da-ja brothers, we define the language $\mathcal{L}_{B^*} = \mathcal{L}_B \cup \{c_d, c_j\}$ and the language $\mathcal{L}_{B^*}^{[\cdot]}$, which is obtained from $\mathcal{L}_{B^*}$ by means of quotational closure. Here, c_d and c_j are constant symbols having 'da' and 'ja' as their respective intended interpretations. The riddle of the da-ja brothers is represented by K^*[5].

Definition 2.6 The set of axioms K^*.
K^* is defined as the following theory in $\mathcal{L}_{B^*}^{[\cdot]}$:

$K_1^* : G(l) \leftrightarrow \neg G(r).$
$K_2^* : (b_1 = b_T \wedge b_2 = b_L) \vee (b_1 = b_L \wedge b_2 = b_T)$
$K_3^* : c_d \neq c_j$
$K_4^* : f_A(b_T, [\sigma]) = c_d \vee f_A(b_T, [\sigma]) = c_j$
$K_5^* : f_A(b_T, [\sigma]) = c_d \leftrightarrow f_A(b_L, [\sigma]) = c_j$
$K_6^* : c_d = c_y \leftrightarrow (\sigma \leftrightarrow f_A(b_T, [\sigma]) = c_d)$
$K_7^* : (c_d = c_y \vee c_d = c_n) \wedge \neg(c_d = c_y \wedge c_d = c_n)$
$K_8^* : (c_j = c_y \vee c_j = c_n) \wedge \neg(c_j = c_y \wedge c_j = c_n)$
$K_9^* : c_d = c_y \leftrightarrow c_j = c_n$ □

Analogous to the fundamental 0-principle for K, we can prove a principle for K^* from which a solution for the riddle of the da-ja brothers easily follows. There is also an analogue to the fundamental 1-principle for K. Interestingly, the last principle allows us to determine the truth-value of sentences without having to talk about the meaning of the words 'da' and 'ja' at all.

Proposition 2.4 The fundamental da-ja 0-principle
Let σ be a sentence of $\mathcal{L}_{B^*}^{[\cdot]}$. We have that:

1. $\vdash_{K^*} \sigma \leftrightarrow f_A(b_1, [c_d = c_y \leftrightarrow (b_1 = b_T \leftrightarrow \sigma)]) = c_d$
2. $\vdash_{K^*} \neg\sigma \leftrightarrow f_A(b_1, [c_d = c_y \leftrightarrow (b_1 = b_T \leftrightarrow \sigma)]) = c_n$

Proof: Left to the reader. □

Proposition 2.5 The fundamental da-ja 1-principle
Let σ be a sentence of $\mathcal{L}_{B^*}^{[\cdot]}$. We have that:

1. $\vdash_{K^*} \sigma \leftrightarrow f_A(b_1, [f_A(b_1, [\sigma]) = c_d]) = c_d$
2. $\vdash_{K^*} \neg\sigma \leftrightarrow f_A(b_1, [f_A(b_1, [\sigma]) = c_d]) = c_j$

Proof: Left to the reader. □

At this point, the reader may wonder how the solutions to the riddles, whose correctness is easily verified, are actually obtained. In the next section, we discuss the method of possible worlds, which is constructive tool to obtain such solutions.

[5]The consistency of K^* easily follows from the consistency of K.

2.3.7 The method of possible worlds

In this section, we present a constructive method, the method of possible worlds, by which solutions to the riddles can be obtained. Let us illustrate the method for the riddle of the yes-no brothers. Here, a riddle-solver knows that the *actual world* is one out of four *possible worlds*[6]. A possible world specifies whether left or right is the good road and also, whether b_1 is the truth speaker or the liar. We refer to the worlds as M_{lT}, M_{lL}, M_{rT} and M_{rL}, where:

- $M_{lT} \vDash G(l) \wedge b_1 = b_T$ $\qquad$ $M_{lL} \vDash G(l) \wedge b_1 = b_L$
- $M_{rT} \vDash G(r) \wedge b_1 = b_T$ $\qquad$ $M_{rL} \vDash G(r) \wedge b_1 = b_L$

The four possible worlds M_{lT}, M_{lL}, M_{rT} and M_{rL} correspond with the first, second, third and fourth column of Table 1 respectively.

$G(l)$	1	1	0	0
$b_1 = b_T$	1	0	1	0
$f_A(b_1, [Q]) = c_y$	1	1	0	0
Q	1	0	0	1
$G(l) \leftrightarrow b_1 = b_T$	1	0	0	1

Table 1

The table is read as follows. An entry of 1 (0) means that the sentence written on the same row as that entry is true (false). So the leftmost column corresponds with M_{lT}; the world in which left is the good road and in which b_1 is the truth-speaker. The third row is made up by me. Whatever question Q we come up with, we want to be able to separate $G(l)$ worlds from non-$G(l)$ worlds by addressing the question Q to the brother we are facing (b_1) in that world. The separation is based on the yes-no answer we get on Q. Hence, whatever question Q we address to b_1, when Q allows us to separate worlds we should get a different answer in $G(l)$ worlds than in non-$G(l)$ worlds. To assure this, I filled the truth table in such a way that the sentence $f_A(b_1, [Q]) = c_y$ is true in $G(l)$ worlds and that it is false in non-$G(l)$ worlds.[7] Fix a column. The entry in the fourth row follows from the entries in the second and third row using the fact that the column represents a model of K. The fourth row tells us that we are after a sentence Q that is true in M_{lT} and M_{rL}, while false in M_{lL} and in M_{rT}. In the fifth row, we construct a sentence that has the same truth value as Q in each possible world. As we construct the sentence from degree 0 sentences, we arrive at a solution for the riddle that is itself of degree 0 . We say that we found the solution '$G(l) \leftrightarrow b_1 = b_T$' using the *possible worlds method*.

The pay-off of the possible worlds method is more vividly illustrated by the riddle of the da-ja brothers. Table 2 has the same rationale as Table 1.

[6]The notion of a possible world and actual world can be made precise in terms of the models that are constructed, in the appendix, to prove the consistency of K. However, we feel that doing so only distracts from the main idea.

[7]Of course reversing the truth /falsity ascription would work equally well.

$G(l)$	1	1	1	1	0	0	0	0
$b_1 = b_T$	1	1	0	0	1	1	0	0
$c_d = c_y$	1	0	1	0	1	0	1	0
$f_A(b_1, [Q]) = c_d$	1	1	1	1	0	0	0	0
$f_A(b_1, [Q]) = c_y$	1	0	1	0	0	1	0	1
Q	1	0	0	1	0	1	1	0
$b_1 = b_T \leftrightarrow G(l)$	1	1	0	0	0	0	1	1
$c_d = c_y \leftrightarrow (b_1 = b_T \leftrightarrow G(l))$	1	0	0	1	0	1	1	0

Table 2

The interpretation of the first three rows is clear. The fourth row of the table is made up by me with the same rationale as the third row of Table 1; now the da-ja answer we receive should allow us to separate $G(l)$ worlds from non-$G(l)$ worlds. The fifth row is a translation of the answer of the addressed person ('da'or 'ja') to Q in terms of the familiar 'yes' and 'no'. The fifth row is a convenient intermediate step to arrive at the sixth row. The sixth row contains the truth value of Q in each world. As in Table 1, this row is obtained by 'backwards engineering'. The seventh row is a convenient intermediate step for arriving at the eight row; a sentence of $\mathcal{L}^{[\cdot]}_{B^*}$ that has the same truth-value in each world as Q. Thus, '$c_d, c_y \leftrightarrow b_1 = b_T \leftrightarrow G(l)$' is the question we are after. In natural language:

'da' means yes iff (you are the truth-speaker iff left is the good road)

2.4 The Hardest Logic Puzzle Ever

2.4.1 The riddle

Now that we know how to solve the riddle of the da-ja brothers, we are ready to discuss the *Hardest Logic Puzzle Ever* ($HLPE$) that was stated in the introduction. As discussed there, Rabern and Rabern observed that there is an ambiguity with respect to instruction (B_3), so that we have to distinguish between a syntactic and semantic version of $HLPE$. Moreover, Uzquiano observed the need to distinguish between a version of $HLPE$ in which True and False have the ability to predict the outcome of the random coin flip occurring in Random's head and one in which they do not have this ability. As the two distinctions are independent, we have four distinct versions of $HLPE$, which are denoted $HLPE^{omn}_{syn}$, $HLPE^{ran}_{syn}$, $HLPE^{omn}_{sem}$ and $HLPE^{ran}_{sem}$. As stated in the introduction, we will not formalize $HLPE^{ran}_{sem}$ for, as will become clear, there is no interest in doing so. Moreover, in the rest of this paper, we will not study the three $HLPE$ variants as presented in the introduction, but rather their 'yes-no' versions. That is, in the $HLPE$ variants that we will actually study, the gods are assumed to answer with 'yes' and 'no' instead of with 'da' and 'ja'. This is not a real restriction, as all our observations also go through for the da-ja setting by modifications similar to those that were involved in Section 2, were the riddle of the da-ja brothers was seen to be a modification of the riddle of the yes-no brothers; we simply gain in economy of presentation without missing any relevant conceptual issue involved in the structure of $HLPE$. As Boolos ([9]) remarks in a footnote (my italics):

The *extra twist* of not knowing which are the gods' words for 'yes' and 'no' is due to the computer scientist John McCarthy.

Indeed, the da-ja feature is nothing but an extra twist which can safely be neglected for our purposes. As we will see, the three-question solutions due to Boolos and Roberts can be obtained in each of the three considered versions of $HLPE$. Their solutions have the following three stage structure.

A Ask a question $\mathbf{Q_1}$ to B the answer to which allows you to identify a non-Random god X (X is either A or C).
B Ask a question $\mathbf{Q_2}$ to X the answer to which allows you to determine whether X is True or False.
C Ask a question $\mathbf{Q_3}$ to X the answer to which allows you to determine the identity of god B. You now know the identity of two gods, so you also know the identity of the third, and you thus solved the riddle.

The solutions of Boolos and Roberts, adapted to suit the yes-no versions of $HLPE$, are as follows:

$\mathbf{Q_{B1}}$: you are True iff A is Random?
$\mathbf{Q_{B2}}$: A is True or A is not True?
$\mathbf{Q_{B3}}$: B is Random?

$\mathbf{Q_{R1}}$: If I asked you if A was Random would you answer 'yes'?
$\mathbf{Q_{R2}}$: If I asked you if you were True would you answer 'yes'?
$\mathbf{Q_{R3}}$: If I asked you if B was Random would you answer 'yes'?

First, we will show how to derive these solutions in a formal presentation of $HLPE^{omn}_{syn}$.

2.4.2 Modeling $HLPE^{omn}_{syn}$: the theory O^{syn}

In this section, we model $HLPE^{omn}_{syn}$ as a theory, called O^{syn}, which is formulated in $\mathcal{L}^{[\cdot]}_G$, the" language of the Gods". $\mathcal{L}^{[\cdot]}_G$ is obtained from the language $\mathcal{L}_G = \{a, b, c, g_T, g_F, g_R, c_y, c_n, f_A\}$ by quotational closure. Here a, b, c are constant symbols for god A, god B and god C and g_T, g_F and g_R are constant symbols for True, False and Random. The constant symbols, c_y and c_n and the binary function symbol f_A are interpreted as in Section 2.3.4. In order to introduce the axioms below, we introduce the following notation. We use σ_{pqr} (where p, q, r are constant symbols of $\mathcal{L}^{[\cdot]}_G$) as shorthand for sentences of $\mathcal{L}^{[\cdot]}_G$ in the following way:

$$\sigma_{pqr} := p = g_T \wedge q = g_F \wedge r = g_R$$

Definition 2.7 The set of axioms O^{syn}
O^{syn} consists of the following axioms / axiom schemes of $\mathcal{L}^{[\cdot]}_G$.
$O^{syn}_1 : \sigma_{abc} \vee \sigma_{acb} \vee \sigma_{bac} \vee \sigma_{bca} \vee \sigma_{cab} \vee \sigma_{cba}$
$O^{syn}_2 : g_R \neq g_T \wedge g_R \neq g_F$
$O^{syn}_3 : c_y \neq c_n$
$O^{syn}_4 : f_A(\mu, [\sigma]) = c_y \vee f_A(\mu, [\sigma]) = c_n \qquad (\mu \in \{g_T, g_F, g_R\})$

$O_5^{syn} : f_A(g_T, [\sigma]) = c_y \leftrightarrow \ f_A(g_F, [\sigma]) = c_n$
$O_6^{syn} : \sigma \leftrightarrow f_A(g_T, [\sigma]) = c_y$ □

We can formulate the $\mathbf{Q_{Bi}}$ and $\mathbf{Q_{Ri}}$ questions of the previous section in the language $\mathcal{L}_G^{[\cdot]}$ and show that these questions constitute a solution for $HLPE$ by derivations in O^{syn}. To do so, it is convenient to have the following two propositions at our disposal.

Proposition 2.6 The fundamental O^{syn} 0-principle
Let σ be a sentence of $\mathcal{L}_G^{[\cdot]}$. Let $\lambda \in \{a, b, c\}$. Then:

1. $\vdash_{O^{syn}} \lambda \neq g_R \rightarrow (\sigma \leftrightarrow f_A(\lambda, [\lambda = g_T \leftrightarrow \sigma]) = c_y)$
2. $\vdash_{O^{syn}} \lambda \neq g_R \rightarrow (\neg\sigma \leftrightarrow f_A(\lambda, [\lambda = g_T \leftrightarrow \sigma]) = c_n)$

Proof: Left to the reader. □

Proposition 2.7 The fundamental O^{syn} 1-principle
Let σ be a sentence of $\mathcal{L}_G^{[\cdot]}$ and let $\lambda \in \{a, b, c\}$. Then:

1. $\vdash_{O^{syn}} \lambda \neq g_R \rightarrow (\sigma \leftrightarrow f_A(\lambda, [f_A(\lambda, [\sigma]) = c_y]) = c_y)$
2. $\vdash_{O^{syn}} \lambda \neq g_R \rightarrow (\neg\sigma \leftrightarrow f_A(\lambda, [f_A(\lambda, [\sigma]) = c_y]) = c_n)$

Proof: Left to the reader. □

We now illustrate how O^{syn} allows us to formalize Boolos' solution. Below, Q_{B1}, Q_{B2} and Q_{B3} are the formalizations of Boolos' questions $\mathbf{Q_{B1}}$, $\mathbf{Q_{B2}}$ and $\mathbf{Q_{B3}}$ that were mentioned in the previous subsection.

A *Identifying a non-random god:*

1. $\vdash_{O^{syn}} f_A(b, [b = g_T \leftrightarrow a = g_R]) = c_y \rightarrow c \neq g_R$
2. $\vdash_{O^{syn}} f_A(b, [b = g_T \leftrightarrow a = g_R]) = c_n \rightarrow a \neq g_R$

This follows from Proposition 2.6 and the fact that b is either g_T, g_F or g_R . Thus, by asking $Q_{B1} \coloneqq b = g_T \leftrightarrow a = g_R$ to b we identify a god (a or c) as non-Random.

B *Determining the identity of the non-Random god.*
Let λ be the god (either a or c) that is known to be non-Random.

1. $\vdash_{O^{syn}} f_A(\lambda, [a = g_T \vee a \neq g_T]) = c_y \rightarrow a = g_T$
2. $\vdash_{O^{syn}} f_A(\lambda, [a = g_T \vee a \neq g_T]) = c_n \rightarrow a = g_F$

This is an immediate consequence from the fact that we address a non Random god and the behavior of True and False. Thus, by asking $Q_{B2} \coloneqq a = g_T \vee a \neq g_T$ to the non-Random god, we reveal the identity (True or False) of the non-Random god.

C *Determining the identity of another god.*
Let λ be the god that is either known to be True or known to be False. Suppose that λ turned out to be True (otherwise interchange c_y and c_n below).

1. $\vdash_{O^{syn}} f_A(\lambda, [b = g_R]) = c_y \rightarrow b = g_R$

2. $\vdash_{O^{syn}} f_A(\lambda, [b = g_R]) = c_n \rightarrow b \neq g_R$

Thus by asking $Q_{B3} \coloneqq b = g_R$ to λ we reveal whether or not b is Random. Clearly, the information obtained by the three questions determines the identity of all three gods.

In a similar way, now using Proposition 2.7 rather than Proposition 2.6, we can show that O^{syn} derives the solution of Roberts. The translations of his questions into $\mathcal{L}_G^{[\cdot]}$ are as follows:

$Q_{R1} \coloneqq f_A(b, [a = g_R]) = c_y$
$Q_{R2} \coloneqq f_A(\lambda, [b = b_T]) = c_y$
$Q_{R3} \coloneqq f_A(\lambda, [b = b_R]) = c_y$

Here, λ is either a or c depending on whether the answer on Q_{R1} revealed that respectively a or c is non-Random. Observe that Roberts questions all have degree 1, while those of Boolos all have degree 0. In section 2.4.4, where we discuss $HLPE_{syn}^{ran}$, we point out some consequences of this difference.

2.4.3 Modeling $HLPE_{sem}^{omn}$: the theory O^{sem}

Remember that a semantic Random behaves according to the following protocol. Whenever we ask a question **Q** to Random, he flips a coin and then, depending on the outcome of the coin-flip, he answers **Q** *truly or falsely*. Elaborating on suggestions in ([43]), we think of the coin-flip as determining the *mental state* in which Random answers **Q**. The mental state, on its turn, determines whether Random answers **Q** correctly or falsely. I take the following three claims as meaning-constitutive for the concept of a mental state. The third statement can be seen as a more precise specification of the semantic protocol, couched in terms of mental states. I shall refer to it as the *mental-state protocol*.

1. There are two mental states, call them T and F.

2. The lying /truth telling behavior of Random on answering a question **Q** is determined by the mental state he has just before he answers **Q**. When that state is T he answers **Q** truthfully, when that state is F he lies.

3. When a question **Q** is addressed to Random, first a coin-flip determines his mental state and then, being in this mental state, Random answers **Q** .

In this section, we will model $HLPE_{sem}^{omn}$ as a theory, called O^{sem}, in the language $\mathcal{L}_{G'}^{[\cdot]}$, the quotational closure of $\mathcal{L}_{G'} = \mathcal{L}_G \cup \{T, F\}$, where T and F are two binary predicate symbols such that $T(g_R)$ reads as 'Random is in mental state T' and $F(g_R)$ is interpreted similarly.

Definition 2.8 The theory O^{sem}
O^{sem} consists of the following axioms / axiom schemes of $\mathcal{L}_{G'}^{[\cdot]}$.

1. All of O^{syn} (For $n = 1 - 6$, set $O_n^{sem} = O_n^{syn}$).

2. The following additional axiom (schemes) specifying Random's behavior:
   O_7^{sem} : $(T(g_R) \vee F(g_R)) \wedge \neg(T(g_R) \wedge F(g_R))$
   O_8^{sem} : $T(g_R) \rightarrow (f_A(g_R, [\sigma]) = f_A(g_T, [\sigma]))$
   O_9^{sem} : $F(g_R) \rightarrow (f_A(g_R, [\sigma]) = f_A(g_F, [\sigma]))$ □

So according to O_7^{sem} Random is either in mental state T or F. Axiom scheme O_8^{sem} tells us that whenever Random is in mental state T, his answer on a question is the same as the answer of True, while axiom scheme O_9^{sem} tells us that whenever Random is in mental state F, his answer is the same as that of False. Due to the presence of the additional axioms $(O_7^{sem}, O_8^{sem}, O_9^{sem})$ we are in fact always addressing a truth speaker or a liar and hence the proviso of Proposition 2.6 and Proposition 2.7, which specified that we are addressing a god that is non-Random, can be removed. That is, the fundamental O^{sem} 0-principle and O^{sem} 1-principle are as follows:

Proposition 2.8 The fundamental O^{sem} 0-principle
Let σ be a sentence of $\mathcal{L}_{G'}^{[\cdot]}$. Let $\lambda \in \{a, b, c\}$. Then:

1. $\vdash_{O^{sem}} \sigma \leftrightarrow f_A(\lambda, [\lambda = g_T \leftrightarrow \sigma]) = c_y$
2. $\vdash_{O^{sem}} \neg\sigma \leftrightarrow f_A(\lambda, [\lambda = g_T \leftrightarrow \sigma]) = c_n$

Proof: Left to the reader. □

Proposition 2.9 The fundamental O^{sem} 1-principle
Let σ be a sentence of $\mathcal{L}_{G'}^{[\cdot]}$ and let $\lambda \in \{a, b, c\}$. Then:

1. $\vdash_{O^{sem}} \sigma \leftrightarrow f_A(\lambda, [f_A(\lambda, [\sigma]) = c_y]) = c_y$
2. $\vdash_{O^{sem}} \neg\sigma \leftrightarrow f_A(\lambda, [f_A(\lambda, [\sigma]) = c_y]) = c_n$

Proof: Left to the reader. □

Observe that we can easily derive Boolos' and Roberts' solution in O^{sem}. Moreover, due to the availability of propositions 2.8 and 2.9, additional and "simpler" solutions are available in O^{sem} than in O^{syn}. In particular, the following three questions, after Rabern and Rabern, constitute a solution in O^{sem} but not in O^{syn}.

1. First ask '$f_A(a, [a = g_T]) = c_y$' to a. From Proposition 2.9, it follows that when a answers with 'yes', a is True, whereas an answer of 'no' indicates that a is not True.
2. *i*) If a answers the first question with 'yes', i.e., if a is True, ask '$b = g_F$' as a follow up question to a, the answer to which allows you to determine the identity of all three gods.
   ii) If a answers the first question with 'no', i.e., if a is not True, ask '$f_A(a, [a = g_F]) = c_y$' as a follow up question to a. From Proposition 2.9, it follows that when a answers with 'yes', a is False, whereas an answer of 'no' indicates that a is not False, from which it follows, as a is not True, that a is Random.

3. Ask '$f_A(a, [b = g_T]) = c_y$' to a as a follow up question to $2ii$). From Proposition 2.9, it follows that a's answer allows you to determine whether or not b is True.

The solution thus essentially involves applications of the fundamental O^{sem} 1-principle. A similar solution can be obtained via the fundamental O^{sem} 0-principle. Note that in these "Rabern and Rabern solutions", all three questions are asked to the same god. This is not the case for the solutions of Boolos and Roberts, for which it is essential that different gods are addressed. A more significant distinction between the two types of solutions is revealed when we compare them on the basis of the *average* number of questions that have to be addressed in order to solve the puzzle. To solve $HLPE^{omn}_{sem}$ via the questions of Rabern and Rabern, we may—if the first question is answered with 'yes'—only need to ask two questions. As the probability that the first question is answered with 'yes' is $\frac{1}{3}$, the average number of questions needed by the Rabern and Rabern solution is $\frac{1}{3} \cdot 2 + \frac{2}{3} \cdot 3 = 2\frac{2}{3}$. In contrast, to solve $HLPE^{omn}_{sem}$ via the questions of Boolos and Roberts, we *always* need to ask three questions.

2.4.4 Modeling $HLPE^{ran}_{syn}$: the theory R^{syn}

In this section, we will model $HLPE^{ran}_{syn}$, in which Random behaves according to the syntactic protocol and in which True and False do not have the ability to predict Random's answers. We will compare the solutions of Boolos and Roberts, which also go through for $HLPE^{ran}_{syn}$, and reveal an interesting dissimilarity in their solutions: for Roberts' solution to go through, it is necessary that the gods know their own answers to the questions, whereas this is not a necessary condition for Boolos' solution. Moreover, we will contrast the three-question solutions of Boolos and Roberts with a two-question solution for $HLPE^{ran}_{syn}$ that is due to Uzquiano. The latter solution explicitly exploits the fact that True and False cannot predict the answers given by Random. As a consequence, True and False will reply to some questions—about Random's answers—with an 'I don't know'. We will see that Uzquiano's solution, just like Boolos' solution, does not depend on the assumption that the gods know their own answers. However, in order to derive Uzquiano's solution for $HLPE^{ran}_{syn}$ formally, we need, in contrast to Boolos' solution, two additional axioms which state explicitly that True and False do not know the answers of Random.

In $HLPE^{omn}_{syn}$, True and False *can* predict the behavior of Random. Formally, this is guaranteed by O^{syn}, as for every sentence σ of $\mathcal{L}^{[\cdot]}_G$, we have that:

$$\vdash_{O^{syn}} f_A(g_R, [\sigma]) = c_y \leftrightarrow f_A(g_T, [f_A(g_R, [\sigma]) = c_y]) = c_y$$

Indeed, O^{syn} ensures that True and False are omniscient with respect to the sentences of $\mathcal{L}^{[\cdot]}_G$. To obtain a formal representation of $HLPE^{ran}_{syn}$, we will restrict the knowledge of True and False. In order to do so, we add a single knowledge predicate, $K(x)$, to our language, that can be read as 'True (False) knows x'. That is, we assume that the knowledge of True and False coincide. Making the answers of the gods True and False dependent on their knowledge, we need to make a decision concerning the answers of those gods on questions on which they do not know the answer. We say that True and False give a 'don't know' answer to such questions, and the constant symbol $c_?$ will be used to model

this answer. The formal language of this section is denoted as $\mathcal{L}_{G^*}^{[\cdot]}$, which is obtained from $\mathcal{L}_G \cup \{c_?, K\}$ by means of quotational closure. $HLPE_{syn}^{ran}$ will be modeled via the theory R^{syn}, which consists of two parts, one part modeling the behavior of the knowledge predicate K (the epistemological knowledge, so to speak) one part modeling the riddle. To describe the behavior of the knowledge predicate, we use the epistemological rule of inference EI, as indicated below. Thus, in addition to the inference rules of first order logic, we employ another rule of inference. We use $\vdash_{R^{syn}} \sigma$ to indicate that σ can be derived from the set of first order sentences R^{syn} *using the inference rules of first order logic and* EI. As usual, the rule EI can only be used in categorical, and not in hypothetical derivations.

Definition 2.9 The theory R^{syn}
The theory R^{syn} consists of axiom schema's E_1, E_2 and E_3, inference rule EI (modeling the epistemological knowledge) and axioms / axiom schema's $R_1 - R_{10}$ (modeling the riddle specific knowledge). In R_5, R_8 and R_{10}, $\mu \in \{g_T, g_F\}$.

$E_1 : K([\sigma]) \to \sigma$
$E_2 : K([K([\sigma]) \to \sigma])$
$E_3 : K([\sigma \to \theta]) \to (K([\sigma]) \to K([\theta]))$
$EI :$ If $\vdash_{R^{syn}} \sigma$ then $\vdash_{R^{syn}} K([\sigma])$

$R_1 : K([\sigma_{abc}]) \vee K([\sigma_{acb}]) \vee K([\sigma_{bac}]) \vee K([\sigma_{bca}]) \vee K([\sigma_{cab}]) \vee K([\sigma_{cba}])$
$R_2 : g_R \neq g_T \wedge g_R \neq g_F$
$R_3 : c_y \neq c_n \wedge c_? \neq c_y \wedge c_? \neq c_n$
$R_4 : f_A(g_R, [\sigma]) = c_y \vee f_A(g_R, [\sigma]) = c_n$
$R_5 : f_A(\mu, [\sigma]) = c_y \vee f_A(\mu, [\sigma]) = c_n \vee f_A(\mu, [\sigma]) = c_?$
$R_6 : K([\sigma]) \leftrightarrow f_A(g_T, [\sigma]) = c_y$
$R_7 : K([\neg\sigma]) \leftrightarrow f_A(g_T, [\sigma]) = c_n$
$R_8 : \neg K([\sigma]) \wedge \neg K([\neg\sigma]) \leftrightarrow f_A(\mu, [\sigma]) = c_?$
$R_9 : f_A(g_T, [\sigma]) = c_y \leftrightarrow f_A(g_F, [\sigma]) = c_n$

$R_{10} : f_A(\mu, [\sigma]) = \nu \leftrightarrow K([f_A(\mu, [\sigma]) = \nu]) \qquad \nu \in \{c_y, c_n, c_?\}$

$R_{11} : \neg K([f_A(g_R, [\sigma]) = c_y]) \wedge \neg K([\neg f_A(g_R, [\sigma]) = c_y])$
$R_{12} : \neg K([f_A(g_R, [\sigma]) = c_n]) \wedge \neg K([\neg f_A(g_R, [\sigma]) = c_n])$ □

E_1 tells us that knowledge is factive, i.e. that knowledge implies truth. E_2 states that it is known (by True and False) that knowledge is factive, while E_3 tells us that the knowledge of True and False is closed under logical consequence. The inference rule EI states that whatever can be inferred from R^{syn} is known by the non-Random gods. Note that, with EI thus stated, E_2 is in fact superfluous as an axiom, as it can be derived from E_1 using EI.[8] The axioms $R_1 - R_{12}$ are self-explanatory. In the next section, R_{10}, R_{11} and R_{12} will be discussed in more detail. Below, we state the analogues of Proposition 2.6 and 2.7 for the theory R^{syn}.

[8] When the inference rule EI and axiom scheme E_1 are added to a theory of arithmetic (PA say), the resulting system becomes inconsistent. This is the *paradox of the Knower*. Roughly, the inconsistency is established via a self-referential sentence and the unavailability of such sentences in our formal framework is what saves us from this paradox.

Proposition 2.10 The fundamental syntactic R^{syn} 0-principle
Let σ be a sentence of $\mathcal{L}^{[\cdot]}_{G^*}$ and let $\lambda \in \{a, b, c\}$. Set $\theta := \lambda = g_T \leftrightarrow \sigma$. Then:

1. $K([\theta]) \vee K([\neg\theta]) \vdash_{R^{syn}} \lambda \neq g_R \rightarrow (\sigma \leftrightarrow f_A(\lambda, [\theta]) = c_y)$
2. $K([\theta]) \vee K([\neg\theta]) \vdash_{R^{syn}} \lambda \neq g_R \rightarrow (\neg\sigma \leftrightarrow f_A(\lambda, [\theta]) = c_n)$

Proof: Left to the reader. □

Proposition 2.11 The fundamental syntactic $HLPE$ 1-principle
Let σ be a sentence of $\mathcal{L}^{[\cdot]}_{G^*}$ and let $\lambda \in \{a, b, c\}$. Set $\kappa := f_A(\lambda, [\sigma]) = c_y$. Then:

1. $K([\kappa]) \vee K([\neg\kappa]) \vdash_{R^{syn}} \lambda \neq g_R \rightarrow (\sigma \leftrightarrow f_A(\lambda, [\kappa]) = c_y)$
2. $K([\kappa]) \vee K([\neg\kappa]) \vdash_{R^{syn}} \lambda \neq g_R \rightarrow (\neg\sigma \leftrightarrow f_A(\lambda, [\kappa]) = c_n)$

Proof: Left to the reader. □

The antecedent of the derivation relation involved in the statements of both propositions can be seen as the condition under which we can determine whether or not σ by asking a single question to a god that is not Random. We shall refer to these antecedents of Proposition 2.10 and 2.11 as the *application condition for the θ-strategy and the κ-strategy* respectively. Let us now turn to a comparison of Boolos and Roberts solution in R^{syn}.

Comparing Boolos and Roberts solutions in R^{syn}

As we saw in Section 3.2, the derivation of the Boolos and Roberts solution in system O^{syn} depended essentially (in step **A**) on an application of Proposition 2.6 respectively Proposition 2.7. The analogues of Proposition 2.6 and 2.7 for the system R^{syn} are Proposition 2.10 and 2.11 respectively and so, the O^{syn} derivations of the Boolos and Roberts solution can be easily translated into R^{syn} derivations, *when the application conditions of the θ and κ strategy are fulfilled with $\sigma := a = g_R$ and $\lambda := b$.* In fact, these conditions are fulfilled. For, as the reader may wish to verify, we have that:

$$\vdash_{R^{syn}} K([b = g_T \leftrightarrow a = g_R]) \vee K([\neg(b = g_T \leftrightarrow a = g_R)]) \tag{2.4}$$

$$\vdash_{R^{syn}} K([f_A(b, [a = g_R]) = c_y]) \vee K([\neg f_A(b, [a = g_R]) = c_y]) \tag{2.5}$$

And so, the solutions of Boolos and Roberts can be obtained in R^{syn} as well. However, as we will show, for the solution of Roberts to go through, R_{10}, which specified that the gods know their own answers, is needed, while this is not the case for the solution of Boolos.

Let R^* denote the system that is obtained by removing R_{10} (and R_{11}, R_{12}) from R^{syn} and by modifying EI accordingly. Then:

$$\vdash_{R^*} K([b = g_T \leftrightarrow a = g_R]) \vee K([\neg(b = g_T \leftrightarrow a = g_R)]) \tag{2.6}$$

$$\nvdash_{R^*} K([f_A(b, [a = g_R]) = c_y]) \vee K([\neg f_A(b, [a = g_R]) = c_y]) \tag{2.7}$$

Thus, Roberts' solution essentially depends on axiom schema R_{10} while Boolos' solution can also be derived in the weaker system R^*. The proof of (2.6) is an

easy derivation that can be obtained from R_1 and the epistemological axioms. It is left to the interested reader. We give a model-theoretic argument for (2.7). Consider the structure $M = \langle D, I\rangle$ with $D = Sen(\mathcal{L}^{[\cdot]}_{G^*}) \cup \{\mathbf{g_T}, \mathbf{g_F}, \mathbf{g_R}, \mathbf{c_y}, \mathbf{c_n}, \mathbf{c}_?\}$. Let $I([\sigma]) = \sigma$ for every $\sigma \in Sen(\mathcal{L}^{[\cdot]}_{G^*})$, $I(a) = \mathbf{g_T}, I(b) = \mathbf{g_F}, I(c) = \mathbf{g_R}$ and let the constants $g_T, g_F, g_R, c_y, c_n, c_?$ be interpreted by their bold-faced correlates. We define $I(K)$ and $I(f_A)$ as follows.

1. $I(f_A)(d_1, d_2) = \mathbf{c}_?$ whenever $(d_1, d_2) \notin \{\mathbf{g_T}, \mathbf{g_F}, \mathbf{g_R}\} \times Sen(\mathcal{L}^{[\cdot]}_{G^*})$
2. $I(f_A)(\mathbf{g_R}, \sigma) = \mathbf{c_n}$
3. If $\vdash_{R^*} \sigma$ then: $\sigma \in I(K)$ and $I(f_A)(\mathbf{g_T}, \sigma) = \mathbf{c_y}$ and $I(f_A)(\mathbf{g_F}, \sigma) = \mathbf{c_n}$
4. If $\vdash_{R^*} \neg\sigma$ then: $\neg\sigma \in I(K)$ and $I(f_A)(\mathbf{g_T}, \sigma) = \mathbf{c_n}$ and $I(f_A)(\mathbf{g_F}, \sigma) = \mathbf{c_y}$
5. If $\sigma_{abc} \vdash \sigma$ then: $\sigma \in I(K)$ and $I(f_A)(\mathbf{g_T}, \sigma) = \mathbf{c_y}$ and $I(f_A)(\mathbf{g_F}, \sigma, n) = \mathbf{c_n}$
6. If $\sigma_{abc} \vdash \neg\sigma$ then: $\neg\sigma \in I(K)$ and $I(f_A)(\mathbf{g_T}, \sigma) = \mathbf{c_n}$ and $I(f_A)(\mathbf{g_F}, \sigma) = \mathbf{c_y}$
7. Except for the sentences of steps 3,4,5,6 no other objects are in $I(K)$. For every $\sigma \notin I(K)$ we have that:$I(f_A)(\mathbf{g_T}, \sigma) = I(f_A)(\mathbf{g_F}, \sigma) = \mathbf{c}_?$

It is easily established that the structure M is a model for R^*. The rationale behind the construction is as follows. In step 3 and 4 we put the sentences in $I(K)$ that are forced upon us by EI and we interpret f_A accordingly. Then, we make sure that the structure M validates axiom R_1. We do so by putting the sentence σ_{abc} in the extension of K, as well as all of its first order consequences and by interpreting f_A accordingly. We put the first order consequences of σ_{abc} in $I(K)$ to validate the epistemological axioms E_1, E_2 and E_3. With respect to equation (2.7) we observe that '$f_A(b, [a = g_R]) = c_y$' is not in $I(K)$, as it it can not be derived in R^* and as it is no first order consequence of σ_{abc}. Similar for '$\neg f_A(b, [a = g_R]) = c_y$'. Hence we have that:

$$M \not\models K([f_A(b, [a = g_R]) = c_y]) \vee K([\neg f_A(b, [a = g_R]) = c_y]) \tag{2.8}$$

Thus, for Roberts solution to go through we explicitly need R_{10}, stating that the gods True and False are self-reflective, in the sense that they know their own answers to the questions they are asked.

Uzquiano's two-question solution for $HLPE^{ran}_{syn}$ in R^{syn}

In $HLPE^{ran}_{syn}$, there are questions, such as 'does Random answer 'snow is white' with 'yes'?' which neither True nor False can answer. As Random—who behaves according to the syntactic protocol—has no problems with answering any question whatsoever, the fact that a god fails to answer a question indicates that he is not Random. Uzquiano cleverly exploits the knowledge restriction of True and False into a two-question solution for $HLPE^{ran}_{syn}$. In this section, we represent this solution in R^{syn}. To do so, we employ the following proposition. As its proof is somewhat more complicated than the proofs of the other propositions in this paper, it is given in the appendix.

Proposition 2.12 Uzquiano's lemma
Let $\lambda, \delta \in \{a, b, c\}$ and set $Q^{\lambda}_{\delta} := f_A(\lambda, [c_y = c_n]) = f_A(\delta, [c_y = c_n])$. Question Q^{λ}_{δ} will be addressed to λ and, as such, asks λ whether his answer to the false sentence '$c_y = c_n$' is the same as the answer of δ to that sentence. Here is *Uzquiano's lemma*:

$$\vdash_{R^{syn}} \delta = g_R \leftrightarrow f_A(\lambda, [Q^{\lambda}_{\delta}]) = c_?$$

Proof: See appendix. □

As the reader can verify in the appendix, the right to left direction of Uzquiano's lemma can be obtained in the theory R^* that was discussed in the previous section, i.e., in R^{syn} minus R_{10} and R_{11}. To obtain the left to right direction however, R_{11} and R_{12} (but not R_{10}) is needed. Now for Uzquiano's solution. First, we ask Q^a_b to a. We can show that:

$$\vdash_{R^{syn}} f_A(a, [Q^a_b]) = c_? \rightarrow b = g_R \tag{2.9}$$

$$\vdash_{R^{syn}} f_A(a, [Q^a_b]) = c_y \rightarrow (a = g_R \vee (a = g_F \wedge b = g_T)) \tag{2.10}$$

$$\vdash_{R^{syn}} f_A(a, [Q^a_b]) = c_n \rightarrow (a = g_R \vee (a = g_T \wedge b = g_F)) \tag{2.11}$$

Equation (2.9) is an instantiation of the right to left direction of Uzquiano's lemma. To prove (2.10) and (2.11), which we also do in the appendix, we need the left to right direction of the lemma. We see that the three possible answers allow us identify a god that is not Random: if Q^a_b is answered with $c_?$, b is Random (and so a is not), otherwise, b is not Random. As a follow up question we ask, with $\lambda \in \{a, b\}$ a god that we now know that is not Random, Q^{λ}_c to λ. We now get that:

$$\vdash_{R^{syn}} f_A(\lambda, [Q^{\lambda}_c]) = c_? \rightarrow c = g_R \tag{2.12}$$

$$\vdash_{R^{syn}} f_A(\lambda, [Q^{\lambda}_c]) = c_y \rightarrow (\lambda = g_F \wedge c = g_T) \tag{2.13}$$

$$\vdash_{R^{syn}} f_A(\lambda, [Q^{\lambda}_c]) = c_n \rightarrow (\lambda = g_T \wedge c = g_F) \tag{2.14}$$

The proof of equations (2.12), (2.13) and (2.14) is similar to the proof of (2.9), (2.10) and (2.11). If Q^{λ}_c is answered with $c_?$, c is Random. But then b is not Random and so the answer to our first question, Q^a_b, was c_y or c_n. If the answer to our first question was c_y, a is False and b is True, whereas if the answer was c_n, a is True and b is False. If Q^{λ}_c is answered with c_y or c_n, we clearly know the identity of all three gods.

Uzquiano thus solves $HLPE^{ran}_{syn}$ by asking only two (non-self-referential) questions. A distinguishing feature of Uzquiano's solution with all other solutions to $HLPE$, is that, when the second question is answered with $c_?$, we have to refer back to the answer we got on our first question to determine the identity of all three gods. Although we presented this "backtracking argument" in our meta-language, it is clear that it can be represented directly in R^{syn} as well. As we will see, the self-referential solution of Uzquiano to $HLPE^{omn}_{syn}$, to be presented in the next section, also involves "backtracking".

2.5 Self-referential solutions to $HLPE$

The formal systems that modeled the various riddles discussed in this paper all have a common feature; self-referential sentences cannot be formulated in those systems and hence self-referential questions cannot be asked to the gods. This is somewhat disappointing, as when it is allowed to ask self-referential questions, $HLPE^{omn}_{sem}$ can be solved by asking only two questions to the gods, as pointed out by Rabern and Rabern. Although their solution does not carry over to $HLPE^{omn}_{syn}$, Uzquiano has shown that this riddle allows for a self-referential two-question solution as well. In this section, we discuss the self-referential solutions of Rabern and Rabern and Uzquiano and sketch a possible formal treatment of those solutions.

2.5.1 The self-referential solution of Rabern and Rabern

The essential feature of the two-question solution of Rabern and Rabern for $HLPE^{omn}_{sem}$ is given by their Tempered Liar Lemma ([43], p110). The lemma can be stated as follows. Suppose that there is some object, which is either red, yellow or green. You have no idea what the actual color is and your task is to find this out by asking questions to True. Clearly, two questions suffice to find out the color of the object. However, Rabern and Rabern point out that the availability of self-referential questions allow you to identify the color of the object (with certainty) by asking a *single* question. To arrive at this conclusion, Rabern and Rabern observe that there are self-referential questions which True cannot answer with 'yes' or 'no' according to his nature (which is to speak truly). An example of such a question is **L**, which is phrased as follows:

$$\mathbf{L}: \text{ is your answer to } \mathbf{L} \text{ 'no'?}$$

If True answers **L** with either 'yes' or 'no', he can be accused of lying. Thus, True cannot answer question **L** in accordance with his nature and we may assume that he remains silent on such unanswerable questions. Whether or not a question is unanswerable for True may depend on the empirical circumstances. In particular, to return to our riddle, it may depend on the color of the object. Let $R(o)$, $Y(o)$ and $G(o)$ be sentences that state that the object is red, yellow and green respectively. Now ask the following question to True.

$$\mathbf{Q}: \text{ (is your answer to } \mathbf{Q} \text{ 'no' and } R(o) \text{) or } Y(o) \text{ ?}$$

Rabern and Rabern argue that the response of True to **Q** (answering 'yes', 'no' or remaining silent) allows you to identify the color of the object. They reason as follows, using reductio ad absurdum.

1. Suppose True answers **Q** with 'yes' and $\neg Y(o)$. Then, as True speaks truly, the left disjunct of **Q** must be true. In particular, this means that True's answer to **Q** is 'no'. Contradiction. Hence, if True answers **Q** with 'yes', the object is yellow.

2. Suppose True answers **Q** with 'no' and $\neg G(o)$. Then, as True speaks truly, **Q** must be false, and in particular, we have $\neg Y(o)$. From $\neg G(o)$ and $\neg Y(o)$, we conclude $R(o)$. As $R(o)$ and as True answered **Q** with 'no', the left disjunct of **Q**, is true, and so it follows that True's answer to

Q must be 'yes'. Contradiction. Hence, if True answers **Q** with 'no', the object is green.

3. Suppose True remains silent on **Q** and $\neg R(o)$. If $Y(o)$, **Q** would be true and so True would answer **Q** with 'yes', which would contradict the assumption that he remains silent. Thus, $\neg Y(o)$. From $\neg R(o)$ and $\neg Y(o)$ it follows that **Q** is false and hence that True must answer **Q** with 'no'. Contradiction. Hence, if True remains silent on **Q**, the object is red.

The proof of the Tempered Liar Lemma certainly sounds convincing. However, the semantical paradoxes testify that reasoning with self-referential sentences is tricky business. And indeed, a closer look reveals that there is something odd about the proof of the Tempered Liar Lemma. For consider the following "proof" of the claim that True never remains silent when he is asked **Q**.

4. Suppose True remains silent on **Q**. If $Y(o)$, **Q** would be true and so True would answer **Q** with 'yes', which would contradict the assumption that he remains silent. Thus, $\neg Y(o)$. From the fact that True remains silent on **Q** it follows that he does not answer **Q** with 'no' and hence the left disjunct of **Q** is false. As $\neg Y(o)$, **Q** itself is false and hence True must answer **Q** with 'no'. Contradiction. Hence, True does not remain silent on **Q**.

If we accept, as I do, the (intuitive) validity of the Tempered Liar Lemma, we have to explain why the reasoning in 3 is legitimate while that in 4 is not. The following two (intuitive) principles allow us to do so.

i σ is true $\Rightarrow$ True answers σ with 'yes' *if in doing so he will not contradict himself.*

ii σ is false $\Rightarrow$ True answers σ with 'no' *if in doing so he will not contradict himself.*

Thus, the "proof" that True does not remain silent on **Q**, i.e., 4, is blocked by principle *ii*. Although the assumption that True remains silent on **Q** renders **Q** false, from the falsity of **Q** we cannot conclude that he answers **Q** with 'no', as in doing so, he contradicts himself. Although an appeal to principle *ii* allows us to understand the Tempered Liar Lemma as intuitively valid, ultimately, we are interested in a formal account of the validity of the Tempered Liar Lemma. In [64], I gave a model theoretic account of the Tempered Liar Lemma according to which it is valid. Below, I sketch the essentials of the approach taken.

The formal framework of this paper forbids the formulation of questions like **L** and **Q**. However, when we remove the restriction that sentences can only be denoted by quotational constants, formal analogues of **L** and **Q** are readily available. For instance, the formal analogue of **L** is obtained by introducing a constant, say θ, and by specifying that it denotes the sentence $f_A(g_T, \theta) = c_n$ (which is then addressed to True). In fact, we can "hard-wire" the information that θ denotes $f_A(g_T, \theta) = c_n$ in our theory by using quotational names:

$$\theta = [f_A(g_T, \theta) = c_n] \tag{2.15}$$

One can think of (2.15) as a definition, by a riddle-solver, of a self-referential question. We may use such definitions to give a formal representation of the

Tempered Liar Lemma. Let $\mathcal{L}^{[\cdot]}$ be the quotational closure of the language $\mathcal{L} = \{R, Y, G, o, \theta_0, f_A, c_y, c_n, c_?\}$. The sentence $\sigma_R \coloneqq (R(o) \wedge \neg Y(o) \wedge \neg G(o))$ abbreviates the sentence which states that the object is red (and not yellow and green) and σ_Y and σ_G are defined similarly. Consider the theory $\mathbf{T}$, consisting of the following two sentences of $\mathcal{L}^{[\cdot]}$.

$\mathbf{T}_0$: $\theta_0 = [(f_A(g_T, \theta_0) = c_n \wedge R(o)) \vee Y(o)]$

$\mathbf{T}_1$: $\sigma_R \vee \sigma_Y \vee \sigma_G$

Thus, $\mathbf{T}_0$ is the formal definition of question $\mathbf{Q}$, while $\mathbf{T}_1$ represents your knowledge about the object before you have asked any question. A *base model* for $\mathcal{L}^{[\cdot]}$ is a sentence structure for $\mathcal{L}^- = \mathcal{L}^{[\cdot]} - \{f_A\}$. Thus, a base model is a classical model in which every sentence σ of $\mathcal{L}^{[\cdot]}$ is denoted by $[\sigma]$ and which validates $\mathbf{T}$. The answering function f_A does not have an interpretation in a base model. There are three relevant (classes of) base models, associated with the three possible colors of the object. The idea of [64] was to combine techniques of Kripke and Gupta to extend a base model for $\mathcal{L}^{[\cdot]}$ to a (partial) model for $\mathcal{L}^{[\cdot]}$ and to formalize the intuitive notion of validity present in the proof of the Tempered Liar Lemma in terms of the extended base models. Very roughly, the conversion of a base model M to its extension $\mathcal{M}$ proceeds as follows.

1. Take a base model M and use Kripke's Strong Kleene minimal fixed point construction to obtain a partial model M' of $\mathcal{L}^{[\cdot]} - \{c_?\}$.

2. Take M' and use a Gupta revision construction to obtain the extended (partial) model $\mathcal{M}$ for $\mathcal{L}^{[\cdot]}$.

The first stage of the construction declares all sentences of $\mathcal{L}^{[\cdot]} - \{c_?\}$ to be true ($\mathbf{t}$), false ($\mathbf{f}$) or ungrounded ($\mathbf{u}$). The second stage reflects on this first stage and ensures that True remains silent on a sentence just in case that sentence is declared to be ungrounded in the first stage. That is, remaining silent is, but answering with 'yes'or 'no' not, modeled as a classical property. Moreover the second stage guarantees that when True is asked whether he remains silent on some sentence, he answers truthfully. The details of the construction do not matter for our purposes. $\mathbf{M}$ denotes the class of all extended models for $\mathcal{L}^{[\cdot]}$, one for each base model. An extended model $\mathcal{M} \in \mathbf{M}$ defines a valuation $V_{\mathcal{M}} : Sen(\mathcal{L}^{[\cdot]}) \to \{\mathbf{t}, \mathbf{u}, \mathbf{f}\}$. We use $V_{\mathcal{M}}$ to define a notion of validity in terms of truth preservation in extended models. That is:

$$\alpha \vDash \beta \Leftrightarrow \forall \mathcal{M} \in \mathbf{M} : V_{\mathcal{M}}(\alpha) = \mathbf{t} \Rightarrow V_{\mathcal{M}}(\beta) = \mathbf{t}$$

For every sentence σ of $\mathcal{L}^{[\cdot]}$, we have that:

$$\vDash f_A(g_T, [\sigma]) = c_? \vee \neg f_A(g_T, [\sigma]) = c_?$$

That is, True either explodes on σ or he does not. This is guaranteed by the second stage of the construction of an extended model. However, we do not have that:

$$\vDash f_A(g_T, [\sigma]) = c_y \vee \neg f_A(g_T, [\sigma]) = c_y$$

This is a consequence of the fact that $f_A(g_T, [\sigma]) = c_y$ may be ungrounded and the behavior of the Strong Kleene schema. With respect to θ_0, the question of interest, we get that:

$$f_A(g_T, \theta_0) = c_y \vDash Y(o) \tag{2.16}$$

$$f_A(g_T, \theta_0) = c_n \vDash G(o) \tag{2.17}$$

$$f_A(g_T, \theta_0) = c_? \vDash R(o) \tag{2.18}$$

These three equations capture the three argument cases of the Tempered Liar Lemma.

Our formal representation of the Tempered Liar Lemma via equations (2.16), (2.17) and (2.18) essentially relies on a meta-linguistic notion of validity. That is, although we have (2.16), we also have:

$$\nvDash f_A(g_T, \theta_0) = c_y \rightarrow Y(o) \tag{2.19}$$

To see this, let $\mathcal{M}$ be an extended model in which the object is red. Then $f_A(g_T, \theta_0) = c_y$ is ungrounded in $\mathcal{M}$ while $Y(o)$ is false. Accordingly, the material implication of (2.19) is ungrounded in $\mathcal{M}$. On the other hand, due to the classical behavior of remaining silent, we do have object languages implications that come close to (2.16), (2.17) and (2.18):

$$\vDash \neg f_A(g_T, \theta_0) = c_? \wedge f_A(g_T, \theta_0) = c_y \rightarrow Y(o) \tag{2.20}$$

$$\vDash \neg f_A(g_T, \theta_0) = c_? \wedge f_A(g_T, \theta_0) = c_n \rightarrow G(o) \tag{2.21}$$

$$\vDash f_A(g_T, \theta_0) = c_? \rightarrow R(o) \tag{2.22}$$

The actual solution that Rabern and Rabern provide for $HLPE^{omn}_{sem}$ consist of embedding a question of the same type as θ_0 in another question in accordance with Proposition 2.3. In our terminology, their first question consists of addressing question θ_R to god a.

$$\theta_R = [f_A(a, [(f_A(a, \theta_R) = c_n \wedge b = g_T) \vee b = g_F]) = c_y]$$

By asking this question to a, we can reveal the identity of god b in one step. Thereafter, another (non self-referential) question suffices to solve the puzzle. This solution for $HLPE^{omn}_{sem}$ can be represented in a similar manner as the Tempered Liar Lemma.

I take it that (2.16), (2.17) and (2.18) come pretty close to a satisfactory formal representation of the Tempered Liar Lemma. On the other hand, I do not want to claim that the formal method used is an adequate way to model the behavior of True in a general setting. Here are two limitations.

A The Kripke-Gupta construction that was developed in [64] explicitly excludes self-referential sentences which are formed with $c_?$. As an example, it is forbidden to define question υ by letting:

$$\upsilon = [f_A(g_T, \upsilon) = c_n \vee f_A(g_T, \upsilon) = c_?]$$

Although neither the self-referential solution of Rabern and Rabern, nor that of Uzquiano, relies on asking such "forbidden questions", ultimately, we want an

account of the answering function of True and False with respect to all kinds of self-referential sentences.
B The Kripke-Gupta approach does not distinguish between ungrounded sentences. For instance, let

$$\tau = [f_A(g_T, \tau) = c_y]$$

If we ask question τ to True, we ask him 'is your answer to this question 'yes'?'. Now True can answer with either 'yes' or 'no'; both count as speaking truly. Yet according to our approach, True will remain silent on τ as τ will be valuated as ungrounded.

To overcome these limitations, and several other, is the topic of further research.

2.5.2 The self-referential solution of Uzquiano

Uzquiano's (non-self-referential) two question solution for $HLPE^{ran}_{syn}$ depended on the fact that, due to a restriction of their knowledge, there are questions that neither True nor False can answer. His two question solution for $HLPE^{omn}_{syn}$ also uses a question that, depending on the circumstances, neither True nor False can answer. This time, it is not the lack of knowledge of True and False that renders the question unanswerable but rather, just as in the Rabern and Rabern solution to $HLPE^{omn}_{sem}$, the self-reference involved in the question. In $HLPE^{omn}_{syn}$, no question is unanswerable for Random however, as he will simply flip a coin when asked any question and will answer with 'yes' or 'no' depending on the outcome of the coin flip. Thus, when we ask a question to a god on which we get no answer, we know that the god under consideration is not Random. Uzquiano exploits this observation to construct a two-question solution for $HLPE^{omn}_{syn}$. We will present his solution using our formal terminology, as it is more convenient to do so than in natural language. First, ask the following question to a:

$$\upsilon_1 = [f_A(a, [(b \neq g_R \wedge a = g_F) \vee (b = g_R \wedge f_A(a, \upsilon_1) = c_n)]) = c_y]$$

Uzquiano argues that:

$$f_A(a, \upsilon_1) = c_? \Leftrightarrow b = g_R$$
$$f_A(a, \upsilon_1) = c_y \Rightarrow (a = g_R) \vee (a = g_F \wedge b = g_T)$$
$$f_A(a, \upsilon_1) = c_n \Rightarrow (a = g_R) \vee (a = g_T \wedge b = g_F)$$

So, if a answers υ_1 with $c_?$, we know that a is not Random, (as b is) whereas if he does not answer with $c_?$, we know that b is not Random. Suppose that we find out that a is not Random. Then ask a question υ_2 :

$$\upsilon_2 = [f_A(a, [(c \neq g_R \wedge a = g_F) \vee (c = g_R \wedge f_A(a, \upsilon_2) = c_n)] = c_y]$$

Uzquiano argues that:

$$f_A(a, \upsilon_2) = c_? \Leftrightarrow c = g_R$$
$$f_A(a, \upsilon_2) = c_y \Rightarrow a = g_F \wedge c = g_T$$
$$f_A(a, \upsilon_2) = c_n \Rightarrow a = g_T \wedge c = g_F$$

These answers allow us to determine the identity of all three gods by an argument that is similar to the one given for Uzquiano's two-question solution for $HLPE^{ran}_{syn}$. Thus, Uzquiano's choice of self-referential questions, together with the "backtracking technique" (see Section 3.4.2) allow him to obtain a two-question solution for $HLPE^{omn}_{syn}$. The techniques that were used to formalize the Tempered Liar Lemma can also be used, *mutatis mutandis*, to formalize Uzquiano's two-question solution.

2.5.3 Concluding remarks

We defined a formal framework for riddles about truth which was applied, in particular, to various versions of $HLPE$. The defined framework allows us to formalize all non self-referential solutions to $HLPE$ that are currently present in the literature. By applying our framework, we revealed some interesting dissimilarities between those solutions that were hidden due to their previous formulation in natural language. In the last section, we sketched a possible way to formalize the self-referential solutions to $HLPE$ as well.

Appendix

I: Consistency proof of K

We begin with the specification of a base-structure M_0 for $\mathcal{L}_B^{[\cdot]}$. The base-structure is no model for K, but it is trivially a model for K_1-K_3.

Definition 2.10 The base-structure M_0.
Let $\mathcal{L}_B^{[\cdot]}$ be the language of Definition 2.4. The sentence-structure $M_0 = \langle D, I_0 \rangle$ for $\mathcal{L}_B^{[\cdot]}$ is defined as follows.

1. $D = \{\mathbf{l}, \mathbf{r}, \mathbf{b_T}, \mathbf{b_L}, \mathbf{y}, \mathbf{n}\} \cup Sen(\mathcal{L}_B^{[\cdot]})$.
2. We divide the interpretation function I_0 into a *fixed* and a *revision* part.

   2a) The *fixed part* of I_0:

   - $I_0(l) = \mathbf{l}$, $I_0(r) = \mathbf{r}$, $I_0(c_y) = \mathbf{y}$, $I_0(c_n) = \mathbf{n}$, $I_0(b_T) = \mathbf{b_T}$, $I_0(b_L) = \mathbf{b_L}$.
   - $I_0([\sigma]) = \sigma$, for all $\sigma \in Sen(\mathcal{L}_B^{[\cdot]})$
   - $I_0(f_A)(d_1, d_2) = \mathbf{n}$ if $d_1 \notin \{\mathbf{b_T}, \mathbf{b_L}\}$ or $d_2 \notin Sen(\mathcal{L}_B^{[\cdot]})$
   - $I_0(b_1) = \mathbf{b_T}$, $I_0(b_2) = \mathbf{b_L}$.
   - $I_0(G) = \{\mathbf{l}\}$

   2c) The *revision part* of I_0:

   - $I_0(f_A)(d_1, d_2) = \mathbf{l}$ if $d_1 \in \{\mathbf{b_T}, \mathbf{b_L}\}$ and $d_2 \in Sen(\mathcal{L}_B^{[\cdot]})$ □

Clearly, the structure M_0 is a model for $K_1 - K_3$ and equally clearly, it is not a model for K. But using a *revision process*, we can find a model for K starting from our base-structure. The revision process is defined as follows.

Definition 2.11 The revision process
With $M_0 = \langle D, I_0 \rangle$ as in Definition 2.10 and with $\alpha \in On$, the sentence-structure $M_\alpha = \langle D, I_\alpha \rangle$ is just like M_0 except for the revision part of $I(f_A)$. Hence, we may write $M_\alpha = M_0 + I_\alpha(f_A)$. The revision part of $I_\alpha(f_A)$ is defined as follows:

- When $\alpha = \beta + 1$ for some $\beta \in On$: $\quad (\sigma \in Sen(\mathcal{L}_B^{[\cdot]}))$
  1. $I_{\beta+1}(f_A)(\mathbf{b_T}, \sigma) = \mathbf{y}$ iff $M_\beta \vDash \sigma$
  2. $I_{\beta+1}(f_A)(\mathbf{b_T}, \sigma) = \mathbf{n}$ iff $M_\beta \vDash \neg\sigma$,
  3. $I_{\beta+1}(f_A)(\mathbf{b_L}, \sigma) = \mathbf{y}$ iff $M_\beta \vDash \neg\sigma$,
  4. $I_{\beta+1}(f_A)(\mathbf{b_L}, \sigma) = \mathbf{n}$ iff $M_\beta \vDash \sigma$,

- When α is a limit ordinal: $\quad (\sigma \in Sen(\mathcal{L}_B^{[\cdot]}), \mathbf{b} \in \{\mathbf{b_L}, \mathbf{b_T}\})$
  1. $I_\alpha(f_A)(\mathbf{b}, \sigma) = \mathbf{y} \Leftrightarrow \exists\gamma\forall\beta(\gamma \leq \beta < \alpha \rightarrow I_\beta(f_A)(d_1, d_2) = \mathbf{y})$
  2. $I_\alpha(f_A)(\mathbf{b}, \sigma) = \mathbf{n} \Leftrightarrow \exists\gamma\forall\beta(\gamma \leq \beta < \alpha \rightarrow I_\beta(f_A)(d_1, d_2) = \mathbf{n})$
  3. $I_\alpha(f_A)(\mathbf{b}, \sigma) = \mathbf{l}$ otherwise. □

To prove that the revision process stabilizes, we need the following lemma that essentially states that when a sentence of degree n is made true by M_{n+1}, it is made true by all models M_α with $\alpha \geq n + 1$.

Lemma 2.1 Preservation of I_{n+2} w.r.t $Sen(\mathcal{L}_B^n)$ higher up.
With $\mathbf{b} \in \{\mathbf{b_T}, \mathbf{b_L}\}, \sigma \in Sen(\mathcal{L}_B^n)$ and $\alpha \in On : \alpha \geq n + 2$ we have that:

$$I_{n+2}(f_A)(\mathbf{b}, \sigma) = I_\alpha(f_A)(\mathbf{b}, \sigma)$$

Proof: As the proof is similar to the proof of Gupta's Main Lemma ([23], p11) we only comment on the structure of the proof, which is as follows. Suppose that the lemma is false. Let n^* be the least natural number for which there exists an ordinal such that the lemma fails. Let α^* be the least such ordinal. Clearly, α^* has to be a successor ordinal, say $\alpha^* = \beta + 1$. Thus, n^* and α^* are such that there exists a $\sigma \in Sen(\mathcal{L}_B^{n^*})$ such that:

$$M_0 + I_{n^*+1}(f_A) \vDash \sigma \quad \& \quad M_0 + I_\beta(f_A) \nvDash \sigma \tag{2.23}$$

Now one constructs an isomorphism between the structures M_{n^*+1} and M_β in the language $\mathcal{L}_B^{n^*}$ contradicting (1) and hence contradicting the supposition that the lemma is false. □

From the lemma it easily follows that the ω^{th} structure of the revision process is a model for K:

Theorem 2.1 $M_\omega \vDash K$.
First, we show that $I_\omega(f_A) = I_{\omega+1}(f_A)$. We only need to show that the revision parts of the interpretation functions coincide. So let $(\mathbf{b}, \sigma) \in \{\mathbf{b_T}, \mathbf{b_L}\} \times Sen(\mathcal{L}_B^{[\cdot]})$. Let the degree of σ equal n. By Lemma 2.1 it follows that $I_{n+2}(f_A)(\mathbf{b}, \sigma) = I_\omega(f_A)(\mathbf{b}, \sigma) = I_{\omega+1}(f_A)(\mathbf{b}, \sigma)$. Hence $I_\omega(f_A) = I_{\omega+1}(f_A)$ and thus $M_\omega = M_{\omega+1}$. It is now clear that M_ω is a model for $K_1 - K_5$. What remains to be shown is that it is a model for K_6.

Pick an arbitrary sentence σ. Either $M_\omega \vDash \sigma$ or $M_\omega \vDash \neg\sigma$. Suppose $M_\omega \vDash \sigma$. Then, by definition of our revision process, $I_{\omega+1}(f_A)(\mathbf{b_T}, \sigma) = \mathbf{y}$ and hence $M_{\omega+1} \vDash f_A(b_T, [\sigma]) = c_y$. So, as $M_\omega = M_{\omega+1}$, we have that $M_\omega \vDash f_A(b_T, [\sigma]) = c_y$ and so $M_\omega \vDash \sigma \leftrightarrow f_A(b_T, [\sigma]) = c_y$. Suppose that $M_\omega \vDash \neg\sigma$. Then, by definition of our revision process, $I_{\omega+1}(f_A)(\mathbf{b_T}, \sigma) = \mathbf{n}$ and hence $M_{\omega+1} \vDash f_A(b_T, [\sigma]) = c_n$. So, as $M_\omega = M_{\omega+1}$, we have that $M_\omega \vDash f_A(b_T, [\sigma]) = c_n$ and so $M_\omega \vDash \neg\sigma \leftrightarrow f_A(b_T, [\sigma]) = c_n$. □

II: Deriving Uzquiano's solution

Proposition 12 Uzquiano's lemma
Let $\lambda, \delta \in \{a, b, c\}$ and set $Q^\lambda_\delta := f_A(\lambda, [c_y = c_n]) = f_A(\delta, [c_y = c_n])$. Question Q^λ_δ will be addressed to λ and, as such, asks λ whether his answer to the false sentence '$c_y = c_n$' is the same as the answer of δ to that sentence. We have that:

$$\vdash_{R^{syn}} \delta = g_R \leftrightarrow f_A(\lambda, [Q^\lambda_\delta]) = c_? \tag{2.24}$$

Proof: Observe that we can assume that $\lambda \in \{g_T, g_F\}$, as Random is guaranteed, by (R_3 and) R_4 not to answer with $c_?$. For sake of concreteness, we show that (2.24) holds for $Q^{g_T}_b$. That is, we will show that:

$$\vdash_{R^{syn}} b = g_R \leftrightarrow f_A(g_T, [Q^{g_T}_b]) = c_?, \tag{2.25}$$

where $Q^{g_T}_b := f_A(g_T, [c_y = c_n]) = f_A(b, [c_y = c_n])$. The generalization of (2.25) to (2.24) can safely be left to the reader. First, we prove the left to right direction of (2.25):

1. $\vdash_{R^{syn}} f_A(g_T, [c_y = c_n]) = c_n$ $(R_3, EI,\ R_7)$
2. $\vdash_{R^{syn}} Q^{g_T}_b \rightarrow f_A(b, [c_y = c_n]) = c_n$ (1)
3. $\vdash_{R^{syn}} K([Q^{g_T}_b \rightarrow f_A(b, [c_y = c_n]) = c_n])$ $(2, EI)$
4. $\vdash_{R^{syn}} K([Q^{g_T}_b]) \rightarrow K([f_A(b, [c_y = c_n]) = c_n])$ $(3, E_3)$
5. $\vdash_{R^{syn}} K([f_A(b, [c_y = c_n]) = c_n]) \rightarrow b \neq g_R$ (R_{12})
6. $\vdash_{R^{syn}} K([Q^{g_T}_b]) \rightarrow b \neq g_R$ $(4, 5)$
7. $\vdash_{R^{syn}} \neg Q^{g_T}_b \rightarrow \neg f_A(b, [c_y = c_n]) = c_n$ $(1, R_3, R_4)$
8. $\vdash_{R^{syn}} K([\neg Q^{g_T}_b]) \rightarrow b \neq g_R$ (7 and steps similar to 3,4,5)
9. $\vdash_{R^{syn}} b = g_R \rightarrow \neg K([Q^{g_T}_b]) \wedge \neg K([\neg Q^{g_T}_b])$ $(5,\ 8)$
10. $\vdash_{R^{syn}} b = g_R \rightarrow f_A(g_T, [Q^{g_T}_b]) = c_?$ $(9, R_8)$

Now we establish the right to left direction of (2.25):

1. $\vdash_{R^{syn}} f_A(g_T, [c_y = c_n]) = c_n \wedge f_A(g_F, [c_y = c_n]) = c_y$ $(R_3, EI,\ R_6,\ R_7)$
2. $\vdash_{R^{syn}} b = g_F \rightarrow \neg Q^{g_T}_b$ $(1, R_3)$
3. $\vdash_{R^{syn}} K([b = g_F \rightarrow \neg Q^{g_T}_b])$ (EI)
4. $\vdash_{R^{syn}} b = g_F \rightarrow K([b = g_F])$ (R_1)

5. $\vdash_{R^{syn}} b = g_F \rightarrow K([\neg Q_b^{g_T}])$ $\quad (3, 4, E_3)$

6. $\vdash_{R^{syn}} b = g_F \rightarrow \neg f_A(g_T, [Q_b^{g_T}]) = c_?$ $\quad (5, R_7, R_3)$

7. $\vdash_{R^{syn}} b = g_T \rightarrow K([Q_b^{g_T}])$ $\quad$ (steps similar to 1-5)

8. $\vdash_{R^{syn}} b = g_T \rightarrow \neg f_A(g_T, [Q_b^{g_T}]) = c_?$ $\quad (6, R_7, R_3)$

9. $\vdash_{R^{syn}} b \neq g_R \rightarrow \neg f_A(g_T, [Q_b^{g_T}]) = c_?$ $\quad (5, 7, R_1)$ □

Observe that our proof of the left to right direction of (2.25), and hence of (2.24), involves axiom schema's R_{11} and R_{12}, whereas this is not the case for the right to left direction. For Uzquiano's solution to go through however, we need both directions of (2.24) and so we cannot do without R_{11} and R_{12}.

Deriving equations (2.10) and (2.11)
We will only establish (2.10), as the proof of (2.11) is similar. That is, we show that

$$\vdash_{R^{syn}} f_A(a, [Q_b^a]) = c_y \rightarrow (a = g_R \vee (a = g_F \wedge b = g_T)) \tag{10}$$

From the left to right direction of Uzquiano's lemma, it follows that:

$$f_A(a, [Q_b^a]) = c_y \vdash_{R^{syn}} b \neq g_R \tag{2.26}$$

Moreover, from R_4, R_6 and R_7, we have that:

$$f_A(a, [Q_b^a]) = c_y \vdash_{R^{syn}} \theta_T \vee \theta_F \vee (a = g_R), \tag{2.27}$$

where $\theta_T \coloneqq (K([Q_b^a]) \wedge a = g_T)$ and $\theta_F \coloneqq (K([\neg Q_b^a]) \wedge a = g_F)$. As knowledge is factive (E_1), we get that:

$$f_A(a, [Q_b^a]) = c_y, \theta_T \vdash_{R^{syn}} Q_b^a \wedge a = g_T \tag{2.28}$$

From (2.26) and (2.28),it follows that the antecedent of (2.28) implies that $b = g_F$. From the definition of Q_b^a, it thus follows that:

$$f_A(a, [Q_b^a]) = c_y, \theta_T \vdash_{R^{syn}} f_A(g_T, [c_n = c_y]) = f_A(g_F, [c_n = c_y]) \tag{2.29}$$

As R^{syn} proves the negation of the consequent of (2.29), we can conclude from (2.27) that:

$$f_A(a, [Q_b^a]) = c_y \vdash_{R^{syn}} \theta_F \vee (a = g_R),$$

from which we obtain (2.10).

Chapter 3

On the Behavior of True and False

3.1 Abstract

Uzquiano [53] showed that the Hardest Logic Puzzle Ever ($HLPE$) (in its amended form due to Rabern and Rabern [43]) has a solution in only two questions. Uzquiano concludes his paper by noting that his solution strategy naturally suggests a harder variation of the puzzle which, as he remarks, he does not know how to solve in two questions. Wheeler and Barahona [56] formulated a three question solution to Uzquiano's puzzle and gave an information theoretic argument to establish that a two question solution for Uzquiano's puzzle does not exist. However, their argument crucially relies on a certain conception of what it means to answer *self-referential* yes-no questions *truly* and *falsely*. We propose an alternative such conception which, as we show, allows one to solve Uzquiano's puzzle in two questions. The solution strategy adopted suggests an even harder variation of Uzquiano's puzzle which, as we will show, can also be solved in two questions. Just as all previous solutions to versions of $HLPE$, our solution is presented informally. The second part of the paper investigates the prospects of formally representing solutions to $HLPE$ by exploiting theories of truth.

3.2 Introduction

Recall Boolos' formulation of the *Hardest Logic Puzzle Ever* ($HLPE$):

> The Puzzle: Three gods A, B and C are called, in some order, True, False, and Random. True always speaks truly, False always speaks falsely, but whether Random speaks truly or falsely is a completely *random* matter. Your task is to determine the identities of A, B, and C by asking three yes-no questions; each question must be put to exactly one god. The gods understand English, but will answer all questions in their own language, in which the words for 'yes' and 'no' are 'da' and 'ja' in some order. *You do not know which word means which.* Before I present the somewhat lengthy solution, let me

give answers to certain questions about the puzzle that occasionally arise:

- (B1) It could be that some god gets asked more than one question (and hence that some god is not asked any question at all).
- (B2) What the second question is, and to which god it is put, may depend on the answer to the first question. (And of course similarly for the third question).
- (B3) Whether Random speaks truly or not should be thought of as depending on the flip of a coin hidden in his brain: if the coin comes down heads, he speaks truly, if tails, falsely.
- (B4) Random will answer 'da' or 'ja' when asked any yes-no question. (Boolos [9, p62])

Rabern and Rabern [43] point out the need to distinguish $HLPE$ as literally formulated by Boolos from a version of $HLPE$ which is closely related to it and, as pointed out by Rabern and Rabern, is more properly called 'the hardest logic puzzle ever'. The distinction between the puzzle as formulated by Boolos—which we call $HLPE_{sem}$, for *semantic* $HLPE$—and the amended puzzle—which we call $HLPE_{syn}$, for *syntactic* $HLPE$—only concerns the way in which Random reacts to questions. Suppose that we address a question to Random. Depending on the version of $HLPE$ under consideration, he reacts as follows:

- $HLPE_{sem}$: Random flips a coin and then, depending on the outcome of the coin-flip, answers the question either truly or falsely.

- $HLPE_{syn}$: Random flips a coin and then, depending on the outcome of the coin-flip, answers the question with either 'da' or 'ja'.

Rabern and Rabern show that $HLPE_{sem}$ allows for a solution (in three questions) which is so simple that it almost trivializes the puzzle. Previous commentators (such as Boolos [9] and Roberts [47]) did not realize the possibility of such a simple solution and Rabern and Rabern plausibly suggest that this is due to the fact that these commentators implicitly assumed that Random worked along the lines of $HLPE_{syn}$. Accordingly, we may regard $HLPE_{syn}$ as a corrected version of $HLPE_{sem}$ which is more properly called 'the hardest logic puzzle ever'.

Besides pointing out the distinction between $HLPE_{sem}$ and $HLPE_{syn}$, Rabern and Rabern come up with a solution to $HLPE_{sem}$ which exploits only two (!) questions. To realize their solution, Rabern and Rabern ask the gods *self-referential questions*, which, as they observe, is not prohibited by Boolos' guidelines. However, their solution does not carry over to $HLPE_{syn}$ and so the question arises whether $HLPE_{syn}$ allows for a two-question solution as well.

Uzquiano [53] shows that $HLPE_{syn}$ has a two-question solution[1]. His solution strategy is inspired by Rabern and Rabern's observation that, given their

[1] Actually, Uzquiano distinguishes two versions of $HLPE_{syn}$ and gives two-question solutions for both versions. The versions differ with respect to the abilities of True and False to predict the answers of Random. The first version assumes that True and False cannot predict Random's answers (which seems reasonable given that Random answers *randomly*), while the

nature, True and False cannot answer all yes-no questions with 'da' and 'ja'. In Uzquiano's framework, True and False are said to *remain silent* on questions that they cannot answer with 'da' or 'ja'. Assuming for simplicity that the gods understand and answer in English, an example of a question on which True must remain silent is given by λ.

$$\lambda : \text{Is it the case that: your answer to } \lambda \text{ is 'no'?}$$

In answering λ with either 'yes' or 'no', True can be accused of lying and so True cannot answer λ "in accordance with his nature". Accordingly, True will remain silent when asked λ. This illustrates that in $HLPE_{syn}$, True and False are thought of as having three reactions to questions; besides answering with 'da' and 'ja' they may also remain silent. However, $HLPE_{syn}$ models Random as a random variable over only two of these reactions: answering with 'da' or answering with 'ja'. As Uzquiano observes, a more natural way to model Random in $HLPE_{syn}$ then, is as a *ternary* random variable, the outcome of which determines whether Random answers 'da', 'ja' or remains silent. $HLPE^2_{syn}$ and $HLPE^3_{syn}$ will be used to denote the original version and Uzquiano's version of $HLPE_{syn}$ respectively. Uzquiano solves $HLPE^2_{syn}$ in two questions, but with respect to $HLPE^3_{syn}$, he remarks that: 'I, for one, do not know how to solve this puzzle in two questions.'

Wheeler and Barahona [56] give a three-question solution to $HLPE^3_{syn}$ and give an information theoretic argument which establishes that $HLPE^3_{syn}$ cannot be solved in less than three questions. Although their argument is certainly correct, it crucially relies on the assumption that there are three distinct ways in which the gods answer yes-no questions. But now consider what happens if we ask the following question to True.

$$\tau : \text{Is it the case that: your answer to } \tau \text{ is 'yes'?}$$

Indeed, just as λ is (when asked to True) an interrogative version of the *Liar*, so τ is (when asked to True) an interrogative version of the *Truthteller*. And just as the Truthteller may be valuated as either true or false, so True can answer τ with either 'yes' or 'no'. However, doing so is, in both cases, completely *arbitrary*. Questions like τ do not have a role to play in previous solutions to $HLPE$, and none of the mentioned papers discusses how True answers such questions. In this paper however, questions like τ will have a crucial role to play: they give rise to a *fourth answer*. Exploiting a four-valued answering repertoire, we will show how to solve $HLPE^3_{syn}$ in two questions.

Our alternative account of how True and False answer yes-no questions makes the arbitrariness of answering τ with either 'yes' or 'no' explicit. According to our account, True gives the following answers to λ and τ:

$$\lambda \text{ can } \textit{neither} \text{ be answered with 'yes' nor with 'no'}$$

$$\tau \text{ can be answered } \textit{both} \text{ with 'yes' and 'no'}$$

second version assume that True and False can predict Random's answers (which seems reasonable as True and False are *omniscient*). Uzquiano's solution to the second version is also a solution to the first version but not vice versa. For our purposes, the distinction does not matter: we give a solution that works for both versions.

There is a clear intuitive sense in which answering λ and τ as such is speaking truly. False will answer the mentioned questions as follows.

λ can *both* be answered with 'yes' and with 'no'

τ can be answered *neither* with 'yes' nor with 'no'

Again, there is a clear intuitive sense in which answering λ and τ as such is speaking falsely. A possible justification of working with a four-valued (in contrast to a three-valued) answering repertoire is that the four (linguistic) answers allow us, even in the presence of self-reference, to respect Boolos' instructions, which state that 'True *always speaks* truly' and 'False *always speaks* falsely'. On the other hand, we can also interpret our two non-standard answers in non-linguistic terms, along the following lines: on questions like λ, the algorithm which describes True's behavior yields no solutions, while on questions like τ it yields two solutions. In such cases, True does not answer with 'yes' or 'no', but its two (non-linguistic) answers reflect, respectively, the lack and abundance of solutions. Although we will work with the linguistic version of our non-standard answers, i.e., with 'both' and 'neither', we will return to the two distinct justifications of a four-valued answering repertoire (cf. Section 3.4.3).

An answering repertoire of four answers naturally suggests an "even harder" variation of the hardest logic puzzle ever, $HLPE^4_{syn}$, in which Random is modeled as a four-valued random variable over the possible answers. As we will see, $HLPE^4_{syn}$ can also be solved in two questions. In fact, we will only show how to solve $HLPE^4_{syn}$ in two questions as our solution to $HLPE^4_{syn}$ is easily seen to solve $HLPE^3_{syn}$ as well[2].

All previous solutions to $HLPE$ are presented (informally) in natural language and our solution to $HLPE^4_{syn}$, as presented in Section 3.3, is no exception. However, given the nature of the gods True and False, one would expect that solutions to $HLPE$ allow for a formal representation that is based on a (formal) theory of truth. In Section 3.4, we explore the prospects of such a formal representation, exploiting (Kripkean) fixed point theories of truth. We will see that, using a restricted formal language, the previous solutions to $HLPE$ as well as the solution put forward in Section 3.3, can be given a formal representation. The formal representations are illuminative as they clearly lay bare the differences between the previous solutions to $HLPE$ and the present one. Although our formalization allows us to represent the (informal) solutions to $HLPE$, nevertheless there are some reasons for not being completely satisfied with it, as will be explained Section 3.4. Section 3.4 concludes by discussing the information theoretic argument of [56], which establishes that (given a three-valued answering repertoire) $HLPE^3_{syn}$ cannot be solved in less than three questions. Section 3.5 concludes the paper.

3.3 Solving the puzzles

3.3.1 Gods who answer with 'yes' and 'no'

In this section, we solve $HLPE^4_{syn}$ under the assumption that the gods speak English: they use 'yes' and 'no' to answer positively and negatively respectively.

[2]Note that $HLPE^3_{syn}$ and $HLPE^4_{syn}$ (deliberately) violate Boolos' instruction ($B4$).

In the next section we give up this simplifying assumption and show how to solve $HLPE^4_{syn}$ itself, in which the gods answer with 'da' and 'ja'.

We use the following abbreviations. A, B and C will be used as in Boolos' guidelines and T, F and R will be used to denote True, False and Random respectively. With x an arbitrary question, $N(x)$ reads as 'your answer to x is 'no", while $Y(x)$ reads as 'your answer to x is 'yes"[3]. Before we state our solution to (the English version of) $HLPE^4_{syn}$, we first briefly comment on the algorithm that gives rise to the answers of True and False. First, True and False calculate how their yes/no answers to a question Q influence the truth-value[4] of Q, in light of which they judge these yes/no answers to be correct ($\checkmark$) or incorrect (X). Exploiting the correctness / incorrectness of their yes-no answers with respect to Q, they then determine which of the four possible answers ('yes', 'no', 'both', 'neither') they give to Q. The process is illustrated by the following table.

Table 3.1: Reactions of True and False

Q(uestion)	Y/N	$\mathcal{V}(Q)$	$\checkmark/\mathsf{X}$	True	False
sw : snow is white	$Y(sw)$	true	$\checkmark$	yes	no
	$N(sw)$	true	X		
sb : snow is black	$Y(sb)$	false	X	no	yes
	$N(sb)$	false	$\checkmark$		
$\lambda : N(\lambda)$	$Y(\lambda)$	false	X	neither	both
	$N(\lambda)$	true	X		
$\tau : Y(\tau)$	$Y(\tau)$	true	$\checkmark$	both	neither
	$N(\tau)$	false	$\checkmark$		

Clearly, the yes-no answers of the gods to sw do not influence its truth-value (which is true). Accordingly, answering sw with 'yes' is correct while answering with 'no' is incorrect. Accordingly, True will answer sw with 'yes' while False answers it with 'no'. The yes-no answers of the gods to λ do influence its truth-value. As illustrated by Table 3.1, answering λ with either 'yes' or 'no' is incorrect. As a consequence, True will answer λ with 'neither', while False will answer with 'both'. The answers to questions sb and τ are explained similarly. In Section 3 we will return to this procedure in more detail. Let us now move forward to our solution to the puzzle.

Our two-question solution has the following structure. First, we ask a question which allows us to identify a god which is not Random. Then, we ask a follow up question to the god which we know not to be Random, and use the answer we get to determine the identity of all three gods.

Finding a god that is not Random
Our first question, α_1, is defined as follows:

$$\alpha_1 : (N(\alpha_1) \text{ and } A = R) \text{ or } (Y(\alpha_1) \text{ and } B = R) \text{ or } C = R$$

Table 3.2 investigates the consequences of answering α_1 with 'yes' or 'no' relative

[3] We could use two place answering predicates and remove the indexical "your". However, as this results in a less streamlined presentation, we chose not to do so.

[4] We treat yes-no questions on par with their associated yes-no statements. That is sloppy, but also very convenient.

to the world under consideration (first column) and reports the reactions of True and False to α_1, which are a function of those consequences as we illustrated above.

Table 3.2: Reactions of True and False on α_1

world	Y/N	$\mathcal{V}(\alpha_1)$	✓/ ✗	True	False
$A = R$	$Y(\alpha_1)$	false	✗	neither	both
	$N(\alpha_1)$	true	✗		
$B = R$	$Y(\alpha_1)$	true	✓	both	neither
	$N(\alpha_1)$	false	✓		
$C = R$	$Y(\alpha_1)$	true	✓	yes	no
	$N(\alpha_1)$	true	✗		

Let's explain the first two rows. When A is Random and α_1 is answered with 'yes', α_1 is false—as all its three disjuncts are—and so answering α_1 with 'yes' is incorrect when A is Random. Similarly, when A is Random and α_1 is answered with 'no', α_1 is true and so when A is Random, answering α_1 with 'no' is incorrect as well. So, when A is Random, True will answer α_1 with 'neither', while False will answer it with 'both'. The other entries in the table are explained similarly. We address α_1 to A and extract the following information from his answers.

Table 3.3: Conclusions based on A's answer to α_1

A's answer	Conclusion 1	Conclusion 2
yes	$(A = T$ and $C = R)$ or $A = R$	$B \neq R$
no	$(A = F$ and $C = R)$ or $A = R$	$B \neq R$
neither	$(A = R$ and $A = T)$ or $(A = F$ and $B = R)$ or $A = R$	$C \neq R$
both	$(A = R$ and $A = F)$ or $(A = T$ and $B = R)$ or $A = R$	$C \neq R$

Conclusion 1 is only an intermediate stage for arriving at Conclusion 2, which, as a function of A's answer to α_1, states which god is not Random. Table 3.3 is, in combination with Table 3.2, self-explanatory.

Determining the identity of A, B and C by a follow up question
By asking question α_1 to A, we either learn that B is not Random or that C is not Random. We assume that we learn that B is not Random, the case where C is not Random being similar. As B is not Random, exactly one of the following four statements is true:

$p_1 :=$ $B = T$ and $A = F$ and $C = R$. $\quad$ $p_2 :=$ $B = T$ and $A = R$ and $C = F$.
$p_3 :=$ $B = F$ and $A = T$ and $C = R$. $\quad$ $p_4 :=$ $B = F$ and $A = R$ and $C = T$.

We will ask B, whom we know not to be Random, question α_2:

$$\alpha_2 : (N(\alpha_2) \text{ and } p_1) \text{ or } (Y(\alpha_2) \text{ and } p_2) \text{ or } p_3$$

Table 3.4 has exactly the same rationale as Table 3.2:

Table 3.4: Reactions of True and False on α_2

world	Y/N	$\mathcal{V}(\alpha_2)$	status Y/N	True	False
p_1	$Y(\alpha_2)$	false	✗	neither	both
	$N(\alpha_2)$	true	✗		
p_2	$Y(\alpha_2)$	true	✓	both	neither
	$N(\alpha_2)$	false	✓		
p_3	$Y(\alpha_2)$	true	✓	yes	no
	$N(\alpha_2)$	true	✗		
p_4	$Y(\alpha_2)$	false	✗	no	yes
	$N(\alpha_2)$	false	✓		

Table 3.5 below, which has exactly the same rationale as Table 3.3, shows that B's answer to α_2 allows us to determine whether p_1, p_2, p_3 or p_4 is the case, which means that B's answer allows us to determine the identity of all three gods.

Table 3.5: Conclusions based on B's answer to α_2

B's answer	Conclusion 1	Conclusion 2
yes	(B = T and p_3) or (B = F and p_4)	p_4
no	(B = T and p_4) or (B = F and p_3)	p_3
neither	(B = T and p_1) or (B = F and p_2)	p_1
both	(B = F and p_2) or (B = F and p_1)	p_2

3.3.2 Gods who answer with 'da' and 'ja'

We will now solve $HLPE^4_{syn}$, in which the gods answer positively and negatively by using, in some order, the words 'da' and 'ja'. The methods of the previous section easily carry over to this slightly more complicated puzzle. Let $M(d, y)$ and $M(d, n)$ abbreviate "da' means 'yes" and "da' means 'no" respectively. Further, with x an arbitrary question, $D(x)$ reads as 'your answer to x is 'da", while $J(x)$ reads as 'your answer to x is 'ja".

Finding a god that is not Random
Our first question, β_1, is defined as follows:

$$\beta_1 : M(d, y) \text{ iff } ((D(\beta_1) \text{ and } A = R) \text{ or } (J(\beta_1) \text{ and } B = R) \text{ or } C = R)$$

In Table 3.6, we investigate the consequences of answering β_1 with 'da' or 'ja' relative to a world in which Random is A, B or C and to a language in which 'da' means either 'yes' or 'no'. The table, depicted below, reports the reactions of True and False to β_1, which are a function of the investigated consequences. Due to our uncertainty with respect to the meaning of 'da' and 'ja', Table 3.6 has 12 (rather than 6) rows. Let us compare row 1 with row 7. The first row tells us that when A is Random and 'da' means 'yes', answering β_1 with 'da' renders β_1 true. As on the first row 'da' means 'yes', answering 'da' to β_1 under the conditions of the first row is correct. Row 7 tells us that, when A is Random and 'da' means 'no', answering 'da' to β_1 renders β_1 false. Accordingly, answering 'da' to β_1 under the conditions of the seventh row is correct. From Table 3.6, it

easily follows that asking β_1 to A allows us to determine the identity of a god which is not Random. Drawing the "conclusion table" associated with Table 3.6 is left to the reader.

Table 3.6: Reactions of True and False on β_1

world	language	D/J	$\mathcal{V}(\beta_1)$	✓/ X	True	False
$A = R$	$M(d,y)$	$D(\beta_1)$	true	✓	both	neither
		$J(\beta_1)$	false	✓		
$B = R$	$M(d,y)$	$D(\beta_1)$	false	X	neither	both
		$J(\beta_1)$	true	X		
$C = R$	$M(d,y)$	$D(\beta_1)$	true	✓	da	ja
		$J(\beta_1)$	true	X		
$A = R$	$M(d,n)$	$D(\beta_1)$	false	✓	both	neither
		$J(\beta_1)$	true	✓		
$B = R$	$M(d,n)$	$D(\beta_1)$	true	X	neither	both
		$J(\beta_1)$	false	X		
$C = R$	$M(d,n)$	$D(\beta_1)$	false	✓	da	ja
		$J(\beta_1)$	false	X		

Determining the identity of A, B and C by a follow up question
By asking question β_1 to A, we either learn that B is not Random or that C is not Random. Again, we assume that we learn that B is not Random, the case where C is not Random being similar. When B is not Random, exactly one of p_1, p_2, p_3 and p_4 is true. As a follow up question to β_1, we will ask β_2 to the non Random god B.

$$\beta_2 : M(d,y) \text{ iff } ((D(\beta_2) \text{ and } p_1) \text{ or } (J(\beta_2) \text{ and } p_2) \text{ or } p_3)$$

Table 3.7, which is depicted below, describes the reactions of True and False to β_2 relative to the world and language under consideration. From Table 3.7, it follows that asking β_2 to B, which is not Random, allows us to determine the identity of all three gods. Drawing the "conclusion table" associated with Table 3.7 is left to the reader.

3.4 Formalizations via Theories of Truth

As noted in the introduction, all the previous solutions to $HLPE$ are presented informally using natural language. In the previous section, we likewise introduced our four-valued conception of True and False informally by showing how it can be applied to solve $HLPE^4_{syn}$. In this section, we discuss the prospects of formally representing the present and previous solutions to $HLPE$. The behavior of the gods True and False in $HLPE$ suggests that a formalization of their behavior can fruitfully be based upon a formal theory of truth. In this section, we follow this suggestion by basing ourselves upon Strong Kleene (Kripkean) fixed point theories of truth. To be sure, there are various theories of truth; we could also work with an account of True and False that is based on say, a revision theory of truth (cf. [24]) or on fixed points that are constructed in accordance with the Supervaluation schema. We choose to work with Strong Kleene theories because such theories are very well-known, easy to present and,

Table 3.7: Reactions of True and False to β_2

world	language	D/J	$\mathcal{V}(\beta_2)$	✓/ ✗	True	False
p_1	$M(d,y)$	$D(\beta_2)$	true	✓	both	neither
		$J(\beta_2)$	false	✓		
p_2	$M(d,y)$	$D(\beta_2)$	false	✗	neither	both
		$J(\beta_2)$	true	✗		
p_3	$M(d,y)$	$D(\beta_2)$	true	✓	da	ja
		$J(\beta_2)$	true	✗		
p_4	$M(d,y)$	$D(\beta_2)$	false	✗	ja	da
		$J(\beta_2)$	false	✓		
p_1	$M(d,n)$	$D(\beta_2)$	false	✓	both	neither
		$J(\beta_2)$	true	✓		
p_2	$M(d,n)$	$D(\beta_2)$	true	✗	neither	both
		$J(\beta_2)$	false	✗		
p_3	$M(d,n)$	$D(\beta_2)$	false	✓	da	ja
		$J(\beta_2)$	false	✗		
p_4	$M(d,n)$	$D(\beta_2)$	true	✗	ja	da
		$J(\beta_2)$	true	✓		

importantly, they allow us to represent the solutions to $HLPE$ in a sense that will be made clear below[5].

In fact, we will not apply our formal modeling to $HLPE$ itself, but rather to *the four roads riddle*, presented in Section 3.4.1. The four roads riddle may be considered as a simplified version of $HLPE$ while containing $HLPE$'s essential features: our formalization of the four roads riddle is easily seen to carry over to (versions of) $HLPE$. The formal language in which we will study the four roads riddle contains a 'yes' and a 'no' predicate, but no "non-standard" answer predicates, such as predicates for 'both', 'neither', 'silence' or what have you. As none of the solutions to $HLPE$ involves questions that are formed using non-standard answer predicates, the expressive limitations of our language do not prevent us from representing these solutions. To be sure, ultimately one wants an account of the behavior of True and False in a more expressive language which does contain non-standard answer predicates. In Section 3.4.3, we will briefly comment on the prospects of such an account.

After presenting the four roads riddle in Section 3.4.1, Section 3.4.2 is concerned with formalizations of the riddle. Section 3.4.3 critically looks back at what has been achieved in Section 3.4.2. Section 3.4.4 discusses the information theoretic argument of [56] that was mentioned in the introduction.

3.4.1 The four roads riddle

3.1.1 The riddle You arrive at a cross roads at which you can head either *north*, *south*, *east* or *west*. You know that only one of the four roads, call it the *good road*, leads to your destination. Unfortunately, you have no clue as to which road is good. However, two gods, call them a and b, are situated at the cross roads. You know that one of these gods is True while the other god is False, but you have no clue as to whether a or b is True. The four roads

[5] Which is not to say that other theories of truth do not allow such representation.

riddle is as follows. Given the circumstances just sketched, can you come up with a single question that, when posed to either one of the gods, allows you to determine which road is good?

3.1.2 The language L_B and its ground models We start out by introducing a restricted formal language in which we will study the four roads riddle. Our basic formal language is a quantifier free[6] predicate language with identity L_B, consisting of the following non-logical vocabulary[7].
Constant symbols:

- a and b, which denote, in some order, **Tr**ue and **Fa**lse.
- g_T and g_F, which denote, respectively, **Tr**ue and **Fa**lse.
- n, w, e, s, which denote, respectively, the **no**rth, **we**st, **ea**st and **so**uth road.
- $\{[\sigma] \mid \sigma \in Sen(L_B)\}$: *quotational constant symbols*[8]; for each $\sigma \in Sen(L_B)$, $[\sigma]$ denotes σ.
- $C = \{c_1, c_2, \ldots, c_n\}$: *non-quotational constant symbols*, which denote (arbitrary) elements of $Sen(L_B)$ and which can be used to define self-referential sentences[9].

Predicate symbols:

- $G(x)$, interpreted as 'x is the good road'.
- $Y(x, y)$ and $N(x, y)$, interpreted as 'the answer of x to y is 'yes" and 'the answer of x to y is 'no" respectively.

A *ground model* $M = (D, I)$ is an interpretation of the "yes/no predicate free fragment of L_B" which respects the intuitive interpretation of L_B that is given above. More precisely, a ground model $M = (D, I)$ is a classical model for $L_B^- = L_B - \{Y, N\}$ which respects the following clauses:

1. $D = \{\mathbf{Tr}, \mathbf{Fa}, \mathbf{no}, \mathbf{ea}, \mathbf{so}, \mathbf{we}\} \cup Sen(L_B)$
2. $I(g_T) = \mathbf{Tr}$, $I(g_F) = \mathbf{Fa}$, $I(n) = \mathbf{no}, I(e) = \mathbf{ea}, I(w) = \mathbf{we}, I(s) = \mathbf{so}$
3. $I([\sigma]) = \sigma$ for all $\sigma \in Sen(L_B)$, $I(c_i) \in Sen(L_B)$ for all $c_i \in C$
4. Either ($I(a) = \mathbf{Tr}$ and $I(b) = \mathbf{Fa}$) or ($I(b) = \mathbf{Tr}$ and $I(a) = \mathbf{Fa}$)
5. Either $I(G) = \{\mathbf{no}\}$ or $I(G) = \{\mathbf{ea}\}$ or $I(G) = \{\mathbf{so}\}$ or $I(G) = \{\mathbf{we}\}$.

[6] We do so for sake of simplicity: the definition of the three- and four-valued answering functions below are easily seen to carry over to quantified languages.

[7] We will use $=, \wedge, \vee, \neg, \rightarrow$ and $\leftrightarrow$ as logical symbolism, the interpretation of which is as expected.

[8] The set of quotational constant symbols has a joint recursive definition together with $Sen(L_B)$, the set of sentences of L_B. The definition of these sets can safely be left to the reader.

[9] For instance, when posed to god a, the sentence $Y(a, c_1)$ may be paraphrased as: 'Is it the case that: your answer to this question is 'yes'?', provided that the denotation of c_1 is $Y(a, c_1)$.

For any ground model M, we will use $\mathcal{C}_M : Sen(L_B^-) \to \{0,1\}$ to denote the (classical) valuation of L_B^- that is induced by M. A ground model fixes all the relevant facts; facts about the world on the one hand and facts about sentential reference on the other. As such, an account of the behavior of True and False owes us an explanation of how True and False answer (arbitrary) L_B sentences relative to a ground model. Below, we are concerned with such explanations.

3.4.2 Formalizations

A three-valued answering function for L_B

Clearly, the predicates $Y(g_T,\cdot)$ and $N(g_T,\cdot)$ bear a close similarity with, respectively, a truth predicate and a falsity predicate. Similarly, the predicates $Y(g_F,\cdot)$ and $N(g_F,\cdot)$ bear a close similarity with, respectively, a falsity predicate and a truth predicate. When we treat our yes /no predicates as truth /falsity predicates in the sense alluded to, Kripke's fixed point techniques, as described in [33], may be readily applied in the present setting. In this section, those techniques will be applied to define a three-valued answering function of True and False with an eye on satisfying the following two desiderata:

A (The construction of) the answering function allows us to represent the previous (three-valued) solutions to $HLPE$.

B The answering function gives the intuitive correct verdict with respect to L_B questions that are not considered in those solutions.

Here we go. By a *(Strong Kleene) fixed point valuation for L_B over a ground model M*, $\mathcal{K}_M : Sen(L_B) \to \{0, \frac{1}{2}, 1\}$, we mean a three-valued valuation of L_B which respects the following five clauses. Below, $\overline{\sigma}$ is an arbitrary constant of L_B (quotational or non-quotational) which denotes $\sigma \in Sen(L_B)$.

1. $\mathcal{K}_M(\sigma) = \mathcal{C}_M(\sigma)$ for all $\sigma \in Sen(L_B^-)$
   $\mathcal{K}_M$ respects the ground model M.

2. $\mathcal{K}_M(Y(g_T,\overline{\sigma})) = \mathcal{K}_M(\sigma)$, $\mathcal{K}_M(N(g_T,\overline{\sigma})) = 1 - \mathcal{K}_M(\sigma)$
   Fixed point condition for $Y(g_T,\cdot)$ and $N(g_T,\cdot)$.

3. $\mathcal{K}_M(Y(g_F,\overline{\sigma})) = 1 - \mathcal{K}_M(\sigma)$, $\mathcal{K}_M(N(g_F,\overline{\sigma})) = \mathcal{K}_M(\sigma)$
   Fixed point condition for $Y(g_F,\cdot)$ and $N(g_F,\cdot)$.

4. $\mathcal{K}_M(Y(t_1,t_2)) = \mathcal{K}_M(N(t_1,t_2)) = 0$, when $I(t_1) \notin \{\mathbf{Tr}, \mathbf{Fa}\}$ or $I(t_2) \notin Sen(L_B)$.
   Only questions receive answers and only gods answer questions.

5. (a) $\mathcal{K}_M(\neg\sigma) = 1 - \mathcal{K}_M(\sigma)$
   (b) $\mathcal{K}_M(\alpha \wedge \beta) = \min\{\mathcal{K}_M(\alpha), \mathcal{K}_M(\beta)\}$
   (c) $\mathcal{K}_M(\alpha \vee \beta) = \max\{\mathcal{K}_M(\alpha), \mathcal{K}_M(\beta)\}$
   $\mathcal{K}_M$ is Strong Kleene.

In general, a ground model M allows us to define various fixed point valuations over it[10]. We could define an answering function for True and False that is based

[10] In the present setting, the number of fixed point valuations over M depends on the denotations of the members of C; if, say, $I(c) = (g_T = g_T)$ for every $c \in C$, there is a unique fixed point valuation over M.

on, say, the minimal fixed point valuation over M or, say, the maximal intrinsic fixed point valuation. As will be clear from the discussion below, these answering functions allow us to represent previous solutions to $HLPE$ (**A**) but, arguably, they do not give the intuitive correct verdict with respect to L_B questions that are not considered by those solutions (**B**). In order to justice to both **A** and **B**, we define the valuation function $\mathcal{K}^\star_M : Sen(L_B) \to \{0, \frac{1}{2}, 1\}$ by quantifying over all Strong Kleene fixed point valuations over M. $\mathcal{K}^\star_M$ is defined as follows, where the quantifiers range over all Strong Kleene fixed point valuations over M.

- $\mathcal{K}^\star_M(\sigma) = 1 \Leftrightarrow \exists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 1 \ \& \ \nexists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 0$
- $\mathcal{K}^\star_M(\sigma) = \frac{1}{2} \Leftrightarrow \nexists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 1 \ \& \ \nexists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 0$
- $\mathcal{K}^\star_M(\sigma) = 0 \Leftrightarrow \exists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 0$

The valuation $\mathcal{K}^\star_M$ is used to define an answering function for True and False as follows.

Answering function based on $\mathcal{K}^\star_M$:

i True (False) answers σ with 'yes' just in case $\mathcal{K}^\star_M(\sigma) = 1$ ($\mathcal{K}^\star_M(\sigma) = 0$).

ii True (False) answers σ with 'no' just in case $\mathcal{K}^\star_M(\sigma) = 0$ ($\mathcal{K}^\star_M(\sigma) = 1$).

iii True and False remain silent on σ just in case $\mathcal{K}^\star_M(\sigma) = \frac{1}{2}$.

Let us first point out why we choose to work with $\mathcal{K}^\star_M$ and not with, say, the minimal or maximal intrinsic fixed point. To do so, consider the following three questions:

$$\theta : \text{Is your answer to } \theta \text{ 'yes' or 'no'?}$$

$$\lambda : \text{Is your answer to } \lambda \text{ 'no'?}$$

$$\tau : \text{Is your answer to } \tau \text{ 'yes'?}$$

To remove the indexical 'your', we assume that the questions are addressed to god a. In order to represent the questions in L_B then, we let θ, λ and τ be non-quotational constants such that $I(\theta) = Y(a, \theta) \vee N(a, \theta)$, $I(\lambda) = N(a, \lambda)$ and $I(\tau) = Y(a, \tau)$. The following table describes how $\mathcal{K}^\star_M$ valuates these questions:

Table 3.8: Values of $\mathcal{K}^\star_M$ for $I(\theta)$, $I(\tau), I(\lambda)$.

World	$\mathcal{K}^\star_M(I(\theta))$	$\mathcal{K}^\star_M(I(\tau))$	$\mathcal{K}^\star_M(I(\lambda))$
$a = g_T$	1	0	$\frac{1}{2}$
$a = g_F$	1	$\frac{1}{2}$	0

Consider question θ. First note that the $\mathcal{K}^\star_M$ account of True and False prescribes that True answer θ with 'yes' and that False answers θ with 'no'. I take

it that this is how it, intuitively, should be[11]. This provides a reason for preferring the $\mathcal{K}^{\star}_{M}$ account of True and False above an account that is based on the minimal fixed point; as θ is *ungrounded*, the minimal fixed point will valuate it as $\frac{1}{2}$, implying that both True and False must remain silent on θ according to the minimal fixed point. To see how θ obtains its $\mathcal{K}^{\star}_{M}$ value, note that $Y(g_T, \cdot)$ and $N(g_F, \cdot)$ are truth predicates in disguise, whereas $Y(g_F, \cdot)$ and $N(g_T, \cdot)$ are disguised falsity predicates. Thus, when posed to True, question θ allows for the alethic paraphrase 'this very sentence is true or false', whereas, when addressed to False, the paraphrase becomes 'this very sentence is false or true'. Clearly then, there is a fixed point in which these sentences are true while there is no fixed point in which they are false; $\mathcal{K}^{\star}_{M}(I(\theta)) = 1$, irrespective of whether we address θ to True or False.

The maximal intrinsic fixed point also valuates θ as 1 and so an account of True and False based on it would prescribe the same answers to θ as the $\mathcal{K}^{\star}_{M}$ account. We prefer the $\mathcal{K}^{\star}_{M}$ account over the account based on the maximal intrinsic fixed point due to the answers that are prescribed to question τ. According to the $\mathcal{K}^{\star}_{M}$ account of True and False, True answers question τ with 'no', whereas False remains silent on τ. Intuitively—as also remarked in [43]—False must indeed remain silent on τ, as he cannot answer it "in accordance with his nature", which is to speak falsely. Although the previous solutions to $HLPE$ do not discuss how True should answer τ, their authors do state that the gods remain silent on a question when they cannot answer that question "in accordance with their nature". But True clearly can answer τ with *either* 'yes' or 'no' "in accordance with his nature"—although doing so is completely arbitrary—and so the question arises how True should answer τ. Now, one may take the arbitrariness of a yes / no answer to τ as a *further* reason for True to remain silent. However, this is not what the authors of previous solutions seem to have in mind[12]. So an account of True and False which prescribes that True answers τ with a yes /no answer seems more in line with the spirit of the previous solutions to $HLPE$. The $\mathcal{K}^{\star}_{M}$ account[13] is such an account, whereas an account based on the maximal intrinsic fixed point is not.

Note that, due to the relations between yes/no predicates and truth /falsity predicates, τ behaves like a Truthteller ('this very sentence is true') when addressed to True while it behaves like a Liar ('this very sentence is false') when addressed to False. As there is a fixed point in which the Truetheller is false, we get that $\mathcal{K}^{\star}_{M}(I(\tau)) = 0$ when a is True. As there is no fixed point in which the Liar is true and no fixed point in which the Liar is false, we get that $\mathcal{K}^{\star}_{M}(I(\tau)) = \frac{1}{2}$ when a is False. The $\mathcal{K}^{\star}_{M}$ valuation of question λ receives a dual explanation.

[11] I take it that question θ reveals an interesting dissimilarity between positively answering a yes-no question and asserting its alethic counterpart: while 'yes' is clearly a *truthful* answer to θ, the ungroundedness of 'this very sentence is true or false' *may* deems its assertion inappropriate. More concretely, answering θ with 'yes' makes it true, while asserting 'this very sentence is true or false' does not render the asserted sentence true.

[12] In [44], Rabern and Rabern comment on the answering function that they had in mind in their published paper: according to this function, True gives a classical (yes /no) answer to questions like τ.

[13] Although the $\mathcal{K}^{\star}_{M}$ account prescribes that True answers τ with 'no', we do not think that there is any further reason to prefer such an account over an account according to which True answers τ with 'yes'. Further, some obvious modifications to $\mathcal{K}^{\star}_{M}$ will yield just such an account.

Putting $\mathcal{K}^{\star}_{M}$ to work

Suppose that—in the setting of the four roads riddle—we (only) want to find out whether or not the north road is good. Asking the question 'is the north road good?' is useless; we do not know whether we address True or False when asking a question. However, a little reflection shows that the following question, when addressed to, say, god a, allows us to find out whether or not the north road is good:

Is your answer to the question 'is the north road good?' 'yes'? (3.1)

The L_B translation of question (3.1) is given by the sentence $Y(a, [G(n)])$. The following table explains, in terms of $\mathcal{K}^{\star}_{M}$, why asking question (3.1) suffices to find out whether or not the north road is good.

Table 3.9: Reactions of a to $Y(a, [G(n)])$.

World	$\mathcal{K}^{\star}_{M}(Y(a, [G(n)]))$	Answer of a
$a = g_T, G(n)$	1	yes
$a = g_T, \neg G(n)$	0	no
$a = g_F, G(n)$	0	yes
$a = g_F, \neg G(n)$	1	no

The table explains that 'yes' indicates that the north road is good and that 'no' indicates that the north road is not good. Question (3.1) is an instance of what is called the Embedded Question Lemma (**EQL**) in [43].

EQL Let E be the function that takes a question Q to the question 'Is your answer to the question 'Q' 'yes'?' When either True or False are asked $E(Q)$, an answer of 'yes' indicates that Q whereas an answer of 'no' indicates that not Q.
Proof: Both a double positive and a double negative make a positive.

Suppose that you addressed question (3.1) to a and that you received 'no' as an answer. So, now you know that either the south, east or west road is good—whereas you don't know whether a is True or False. Hence, we are left with the "three roads riddle". Next, we will show how to solve the three roads riddle via a single question, ρ, that is similar to the (crucial) questions that are exploited by previous (informal) self-referential solutions to $HLPE$. In the spirit of those solutions, we define ρ by referring to ρ in the argument place of the embedding function E of the **EQL**:

ρ : E((Is your answer 'no' to ρ and the south road is good) or the west road is good)

Questions like ρ, which refer to themselves in the argument place of the embedding function E, we call *self-embedded questions*. Note that the solution to $HLPE^{4}_{syn}$ that was given in the previous section does *not* rely on self-embedded questions. We'll return to this observation in Section 3.4.3. In order to explain why the answer to ρ allows us to find out which of the three roads is good, it is convenient to first translate it into L_B. To do so, we let ρ be a non-quotational

constant whose denotation is as follows:

$$Y(a, [(N(a, \rho) \wedge G(s)) \vee G(w)])$$

Here is an intuitive explanation of why ρ does the job. If west is the good road, the embedded question, i.e., $(N(a, \rho) \wedge G(s)) \vee G(w)$, will be true. Hence, True will answer the embedded question with 'yes' and False will answer it with 'no'. Thus, when asked whether they answer the embedded question with 'yes', i.e., when asked ρ, True and False will both answer with 'yes'. Similar reasoning shows that if east is the good road, True and False will both answer with 'no'. Finally, when south is the good road, question ρ reduces to a question which has the same answerhood conditions as the self-embedded question ρ_1:

ρ_1: Is your answer 'yes' to the question of whether your answer to ρ_1 is 'no'?

As observed by [53], neither True nor False can answer ρ_1 in accordance with his nature; they must remain silent on ρ. Similarly, when south is the good road, both True and False must remain silent on ρ_1. The following table—whose construction can safely be left to the reader—shows that our $\mathcal{K}^{\star}_{M}$ based answering function yields the same verdict with respect to the answers of True and False to $I(\rho) = Y(a, [(N(a, \rho) \wedge G(s)) \vee G(w)])$.

Table 3.10: Reactions of a to $I(\rho)$.

World	$\mathcal{K}^{\star}_{M}(I(\rho))$	Answer of a
$a = g_T, G(w)$	1	yes
$a = g_T, G(e)$	0	no
$a = g_T, G(s)$	$\frac{1}{2}$	silence
$a = g_F, G(w)$	0	yes
$a = g_F, G(e)$	1	no
$a = g_F, G(s)$	$\frac{1}{2}$	silence

Table 3.10 reveals the sense in which the $\mathcal{K}^{\star}_{M}$ account of True and False allows us to give a formal representation of the informal solution to the "three roads riddle". The principles at work in the solution to the "three roads riddle" are similar to the principles at work in the previous self-referential solutions to $HLPE$. Accordingly, the $\mathcal{K}^{\star}_{M}$ account of True and False can be used to represent these solutions as well.

Putting a four-valued answering function for L_B to work

We start by defining a four-valued valuation function of L_B, $\mathcal{K}^{\bullet}_{M} : Sen(L_B) \rightarrow \{0, +, -, 1\}$, in a similar manner as we defined $\mathcal{K}^{\star}_{M}$, i.e., by quantifying over all Strong Kleene fixed point valuations. We then define a four-valued answering function for True and False based on $\mathcal{K}^{\bullet}_{M}$ and show how it can be invoked to give a formal representation of a solution to the four roads riddle. The principles at work in our solution to the four roads riddle are similar to the principles at work in our solution to $HLPE^{4}_{syn}$ that was presented in Section 3.3. Hence, the $\mathcal{K}^{\bullet}_{M}$ account can also be used to give a formal representation of our solution to $HLPE^{4}_{syn}$.

Here is the definition of $\mathcal{K}^{\bullet}_{M}$:

- $\mathcal{K}^{\bullet}_{M}(\sigma) = 1 \Leftrightarrow \exists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 1 \;\&\; \nexists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 0$
- $\mathcal{K}^{\bullet}_{M}(\sigma) = - \Leftrightarrow \nexists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 1 \;\&\; \nexists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 0$
- $\mathcal{K}^{\bullet}_{M}(\sigma) = + \Leftrightarrow \exists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 1 \;\&\; \exists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 0$
- $\mathcal{K}^{\bullet}_{M}(\sigma) = 0 \Leftrightarrow \nexists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 1 \;\&\; \exists \mathcal{K}_M : \mathcal{K}_M(\sigma) = 0$

The following table illustrates how $\mathcal{K}^{\bullet}_{M}$ valuates questions θ, τ and λ, i.e., here is the $\mathcal{K}^{\bullet}_{M}$ version of Table 3.7.

Table 3.11: Values of $\mathcal{K}^{\bullet}_{M}$ for $I(\theta)$, $I(\tau), I(\lambda)$.

World	$\mathcal{K}^{\bullet}_{M}(I(\theta))$	$\mathcal{K}^{\bullet}_{M}(I(\tau))$	$\mathcal{K}^{\bullet}_{M}(I(\lambda))$
$a = g_T$	1	+	−
$a = g_F$	1	−	+

According to the four-valued conception of True and False put forward in this paper, True answers τ with 'both', whereas False answers τ with 'neither'. Similarly, according to this conception, True answers λ with 'neither', whereas False answers λ with 'both'. This suggests the following answering function:

Answering function based on $\mathcal{K}^{\bullet}_{M}$:

i True (False) answers σ with 'yes' just in case $\mathcal{K}^{\bullet}_{M}(\sigma) = 1$ ($\mathcal{K}^{\bullet}_{M}(\sigma) = 0$).

ii True (False) answers σ with 'no' just in case $\mathcal{K}^{\bullet}_{M}(\sigma) = 0$ ($\mathcal{K}^{\bullet}_{M}(\sigma) = 1$).

iii True and False answer σ with 'neither' just in case $\mathcal{K}^{\bullet}_{M}(\sigma) = -$.

iv True and False answer σ with 'both' just in case $\mathcal{K}^{\bullet}_{M}(\sigma) = +$.

Suppose that that we want to solve the four roads riddle. It is not hard to calculate that, using the methods of Section 3.3, question γ does the job, where γ denotes:

$$a = g_T \leftrightarrow (N(a,\gamma) \wedge G(n)) \vee (Y(a,\gamma) \wedge G(s)) \vee G(w)$$

By applying the methods of Section 3.3—which can safely be left to the reader—we see that an answer of 'yes' indicates that west is the good road, 'no' indicates that east is good, 'neither' indicates that north is good and 'both' indicates that south is good. The $\mathcal{K}^{\bullet}_{M}$ based answering function True and False yields exactly the same verdicts. To see why, we will consider (only) the case where the north road is good. So, suppose that the north road is good and that you address γ to True. Under these circumstances, the answerhood conditions of γ are equivalent to the answerhood conditions of the following question:

$$\gamma_1 : g_T = g_T \leftrightarrow N(g_T, \gamma_1)$$

As the left-hand side of γ_1 is true, γ_1 is true just in case its right-hand side, i.e., $N(g_T, \gamma_1)$ is true. But $N(g_T, \cdot)$ functions as a falsity predicate and so, $N(g_T, \gamma_1)$ is true just in case γ_1 is false; we get that γ_1 is true just in case it is false. Hence, γ_1 is *paradoxical*. In other words, when a is True and the north road is good, we get that $\mathcal{K}^{\bullet}_{M}(I(\gamma_1) = -$ and so True replies γ_1 with 'neither'.

Now suppose that the north road is good and that you address γ to False. Under these circumstances, the answerhood conditions of γ are equivalent to the answerhood conditions of the following question:

$$\gamma_2 : g_F = g_T \leftrightarrow N(g_F, \gamma_1)$$

As the left-hand side of γ_2 is false, γ_2 is true just in case its right-hand side, i.e., $N(g_F, \gamma_2)$ is false. But $N(g_F, \cdot)$ functions as a truth predicate and so, $N(g_F, \gamma_2)$ is false just in case γ_2 is false; we get that γ_2 is true just in case it is false. Hence, γ_2 is *paradoxical*. In other words, when a is False and the north road is good, we get that $\mathcal{K}^{\bullet}_{M}(I(\gamma_1) = -$ and so False replies γ_2 with 'neither'. So, when the north road is good, both True and False reply to γ with 'neither'. The other three cases are reasoned out similarly and so the answers of True (and False) to γ as obtained according to the method of Section 3.3 are the same as the answers that are obtained via the $\mathcal{K}^{\bullet}_{M}$ based answering function. Similarly, one can show that the solution to $HLPE^{4}_{syn}$ that was presented in Section 3.3, allows for a formal representation using our $\mathcal{K}^{\bullet}_{M}$ based answering function.

Note that our solution to the four roads riddle, i.e., question γ, is not a *self-embedded question*. Likewise, none of the questions discussed in Section 3.3 are self-embedded. In contrast, our solution to the three roads riddle, i.e., question ρ, is a self-embedded question and so are the (crucial) questions invoked in previous self-referential solutions to $HLPE$. In the next section, we will explain, amongst others, in which sense self-embedded questions give rise to a problem for the intuitive interpretation of the answers 'both' and 'neither' that was sketched in the introduction.

3.4.3 Critical Remarks on Formalizations

'Both', 'neither' and self-embedded questions

In Section 3.4.2, we discussed the self-embedded question ρ_1, whose L_B translation is as follows:

$$\rho_1 : Y(a, [N(a, \rho_1)])$$

Due to the self-embedding that is present in ρ_1, it is not clear how the methods of Section 3.3 should be applied to calculate the answers of True and False to it; a yes/no answer to ρ_1 does not (immediately) render the statement true or false. However, it is clear how $\mathcal{K}^{\bullet}_{M}$ valuates ρ_1. Exploiting the similarity between yes /no predicates and truth /falsity predicates, we see that, when addressed to True, ρ_1 may be paraphrased in alethic terms as 'it is true that this very sentence is false'. When addressed to False, the paraphrase becomes 'it is false that this very sentence is true'. Clearly then, we have that $\mathcal{K}^{\bullet}_{M}(I(\rho_1)) = -$, irrespective of whether we address ρ_1 to True or False. But this means that both True and False will answer ρ_1 by replying 'ρ_1 can *neither* be answered with 'yes' or 'no''. Observe that this $\mathcal{K}^{\bullet}_{M}$ prescription is at odds with our original interpretation of the answers 'neither' and 'both', according to which True, in answering 'neither' to question λ speaks truly, whereas False, in answering λ with 'both' speaks falsely. Indeed, as ρ_1 can neither (on pain of a self-contradiction) be answered with 'yes' or 'no', False, in replying with 'neither' cannot be said to answer ρ falsely.

So, although the method of Section 3.3 does not prescribe how the answers to ρ_1 should be calculated, the answers to ρ_1 that are prescribed by $\mathcal{K}^{\bullet}_M$ do not fit in with intended interpretation of 'both' and 'neither'. Thus, two options suggest themselves, which are associated with two distinct conceptions of True and False:

1. Stick to the intended interpretation of 'both' and 'neither' and extend the method of Section 3.3 such that it becomes applicable to self-embedded questions and such that, in particular, the answer given by False to ρ_1 is 'both'. This option is naturally associated with an *informative conception* of True and False in which, in answering our questions, they intend to convey information. For instance, in answering a Liar question λ with 'neither' True intends to convey the information that he can't answer λ with 'yes' or 'no'.

2. Use the $\mathcal{K}^{\bullet}_M$ account of True and False and give up the interpretation of 'both' and 'neither'. For instance, say that if $\mathcal{K}^{\bullet}_M(\sigma) = -$, True and False reply to σ with an explosion, while if $\mathcal{K}^{\bullet}_M(\sigma) = +$, they remain silent. On this account, the non-linguistic actions of exploding and remaining silent have their origin in the "paradoxality" and the "arbitrariness" of the possible yes/ no answers. Being non-linguistic actions, exploding and remaining silent are not to be evaluated in terms of 'speaking truly' and 'speaking falsely'. This option is naturally associated with an *algorithmic conception* of True and False in which, in answering our questions, they do not intend to convey any information, but rather, they follow an algorithm. Explosions and silences, on this conception, are best thought of as two distinct ways in which the algorithm can fail, due to the non-existence of solutions (paradoxality) and the abundance of solutions (arbitrariness) respectively. On the algorithmic conception of True and False, explosions and silences are not genuine *answers*, but rather, states that the gods end up in due to their processing of certain questions.

In a sense, it is more natural to speak of the failure of an algorithm due to the lack of solutions (paradoxality) than due to the lack of the abundance of solutions (arbitrariness). As such, the algorithmic conception of True and False is, arguably, more naturally associated with the 3-valued account of True and False that is underlying the previous solutions to $HLPE$[14].

I take it that option 1 is preferable; I take it that an account of True and False according to which these gods can be understood as always speaking, respectively, truly and falsely, is preferable over an account on which they sometimes do not speak at all. Such an account simply seems to do more justice to Boolos' remarks that 'True *always* speaks truly' and 'False *always* speaks falsely'. To be sure, Boolos may not have envisioned the possibility to ask self-referential questions. Then again, I take it that an account of True and False which manages to respect Boolos' instructions, even in the presence of self-reference, is preferable to an account which does not.

[14] As pointed out by an anonymous referee, it can be argued that a 3-valued algorithmic conception of True and False does not require that we introduce a predicate in our language that represents failures of the algorithm as such failures do not belong to the language of the gods. In other words, it can be argued that the problem of *expressive completeness* (see below) does not arise on a 3-valued algorithmic conception of True and False.

Although it is beyond the scope of this paper to carry out option 1 in a rigorous way, here is a hint of how one may proceed. The crucial aspect of the method of Section 3.3 was that True and False calculate how their yes/no answers to a question σ influence the truth-value of σ, in light of which they judge these yes/no answers to be correct / incorrect. Based on those judgements, they then decide which answer they actually give to σ. So according to the method of Section 3.3, the patterns of reasoning of True and False leading up to the correct / incorrect judgement are exactly the same; they only differ in how these judgements are converted into answers. In particular, with respect to questions like λ and τ, True and False find exactly the same judgements. In a sense, this means that False first calculates whether answering with yes /no is *objectively* correct /incorrect and then decides to lie about these judgements. This idea, of False first calculating whether a yes/no answer is *objectively* correct and *then* lying about his findings, can be put to work in extending the method of Section 3.3 to bear on self-embedded questions. Here is a table which will be used to explain how True (and False) calculate their answer to ρ_1 in this manner.

Table 3.12: Reactions of True and False on ρ_1

Y/N	$\mathcal{V}(N(a,\rho_1))$	$\mathcal{V}(Y(a,[N(a,\rho_1)]))$	✓/ X	True	False
$Y(a,\rho_1)$	false	false	X	neither	both
$N(a,\rho_1)$	true	true	X		

Let us explain. The first row supposes that ρ_1 is answered with 'yes' (by a). As a consequence, the embedded question '$N(a,\rho_1)$' is false, as indicated in the second column. But this means that the *objectively correct* answer to the embedded question is 'no'. Accordingly, $Y(a,[N(a,\rho_1)])$—which may here be thought of as 'the answer that *should* be given to '$N(a,\rho_1)$' is 'yes"—is false, as indicated in the third column. Hence, as $I(\rho_1) = Y(a,[N(a,\rho_1)])$, answering 'yes' to ρ_1 is incorrect. The second row receives a similar explanation. Accordingly, True answers ρ_1 with 'neither' while False answers with 'both'. So according to the envisioned method for processing self-embedded questions, False finds out whether answering with yes/no is objectively correct—and not, say "correct for False"—and lies about his findings. Using exactly the same principles, the answer to "deeper" embedded questions, such as ρ_2, can be calculated.

$$\rho_2 : Y(a,[N(a,[Y(a,\rho_2)])])$$

Although these remarks on self-embedded questions do not define a rigorous algorithm for calculating the answers of True and False, I take it that they illustrate that, despite our possibility to ask self-embedded questions, there are hopes for developing a formal account of True and False according to which they can be understood in accordance with Boolos' instructions. However, self-embedded questions aside, a satisfying formal account of True and False faces more issues that have to be resolved. Below we discusses two such issues.

Expressive incompleteness

As noted before, none of the solutions to $HLPE$ exploits *non-standard questions*, i.e., questions that are formed with "non-standard answer predicates". In particular, L_B only contains answer predicates associated with 'yes' and 'no'

and, in that sense, L_B may be called *expressive incomplete*. Although none of the solutions exploits non-standard questions, it seems reasonable to ask how True and False answer such questions. For instance, how do True and False answer questions like:

$$\mu_1 : \text{Is your answer to } \mu_1 \text{ 'no' or 'neither'?}$$

μ_2 : Do you answer 'yes' to the question of whether you answer 'neither' to μ_2?

Theories of truth typically do not contain predicates associated with the non-classical semantic values they employ in their meta-language. As such, we cannot expect much guidance from theories of truth in developing a formal account of how True and False answer non-standard questions. However, the methods of Section 3.3 do give us some guidance here. Using '$NE(x, y)$' to abbreviate x answers y with 'neither', we can translate questions μ_1 and μ_2 by letting $I(\mu_1) = N(a, \mu_1) \vee NE(\mu_1)$ and $I(\mu_2) = Y(a, [NE(a, \mu_2)])$. Here are the tables which explain how True and False answer μ_1 and μ_2:

Table 3.13: Reactions of True and False on μ_1

Y/N	$\mathcal{V}(N(a,\mu_1) \vee NE(\mu_1))$	✓/ ✗	True	False
$Y(a, \mu_1)$	false	✗	neither	both
$N(a, \mu_1)$	true	✗		

Table 3.13 is self-explanatory. Note that, as True answers μ_1 with 'neither', μ_1 is true. Still, True does not answer μ_1 with 'yes' as in doing so he can be accused of lying. This situation with respect to μ_1—despite being true not being answered with 'yes' by True—has a clear rationale in terms of truthfully answering yes-no questions. However, it also points to a further[15] dissimilarity between yes-no questions and their alethic counterparts. For, consider μ_1', which is the alethic counterpart of μ_1:

$$\mu_1' : \text{ Sentence } \mu_1' \text{ is false or (neither true nor false)}$$

In treating μ_1 and μ_1' alike, we are bound to conclude that μ_1' is 'neither true nor false'. But this exactly *what* μ_1' *says*, and so μ_1' is true and so *not* 'neither true nor false'. In sum, accepting that μ_1' is 'neither true nor false' seems to be tantamount to accepting a contradiction, while accepting that True answers μ_1 with 'neither' has a clear rationale in the (assumed) nature of True.

Table 3.14 below explains how question μ_2, which is a self-embedded question, is answered. Table 3.14 is to be understood along familiar lines. On the first line of Table 3.14, the consequences of answering with 'yes' are considered. Answering μ_1 with 'yes' renders $NE(a, \mu_2)$ false, which ensures that the correct answer to $NE(a, \mu_2)$ is 'no'. Accordingly, $Y(a, [NE(a, \mu_2)]$ is false and so answering μ_2 with 'yes' is incorrect.

Table 3.14: Reactions of True and False on μ_2

Y/N	$\mathcal{V}(NE(a,\mu_2))$	$\mathcal{V}(Y(a,[NE(a,\mu_2)]))$	✓/ ✗	True	False
$Y(a, \mu_2)$	false	false	✗	no	yes
$N(a, \mu_2)$	false	false	✓		

[15] For further dissimilarities, see footnote 11.

Tokens or Types?

Consider the following two questions and suppose that we both address them to True.

$$\lambda: \text{ Is your answer to } \lambda \text{ 'no'?}$$

$$\lambda_1: \text{ Is your answer to } \lambda \text{ 'no'?}$$

Question λ is familiar: True answers it with 'neither'. Question λ_1 asks whether True answers question λ with 'no'. As True answers λ with 'neither' (hence, *not* with 'no'), the truthful answer to λ_1 is 'no'. Hence, True should answer λ_1 with 'no'. Or so it seems. Yet if we base our account of True and False on, say $\mathcal{K}^{\bullet}_{M}$, we get different predictions. Fixed point theories of truth satisfy what is called the *intersubstitutability of truth*[16]. As a consequence, $Y(g_T, \overline{\sigma})$ and σ have the same semantic value according to $\mathcal{K}^{\bullet}_{M}$, whenever $\overline{\sigma}$ denotes σ. In particular then, we have that $\mathcal{K}^{\bullet}_{M}(Y(g_T, \lambda)) = \mathcal{K}^{\bullet}_{M}(Y(g_T, \lambda_1)) = -$ and so according to the $\mathcal{K}^{\bullet}_{M}$ account, True will answer both λ and λ_1 with 'neither'.

It is often argued that theories of truth *should* satisfy the intersubstitutability of truth, cf. Field [15] or Beall [5]. In a nutshell, the argument is that if truth does not satisfy the intersubstitutability property, it can not play its stereotypical role of serving as a device of generalization. Let's accept this argument pertaining to theories that describe the behavior of a truth (and falsity) predicate. Does the argument carry over to a theory that describes the behavior of a 'answers with *yes*' (and 'answers with *no*') predicate? Not necessarily. For one thing, it is not clear that the stereotypical role of a 'answers with *yes*' predicate is to serve as a device of generalization and so the typical argument for the intersubstitutablity breaks down: although the analogy between true /falsity predicates and yes/no predicates is close, it is not perfect, as we also noted in the previous subsection. I take the intuitive reasoning with respect to λ and λ_1 that was given above convincing and I do not see why the intersubstitutablity of *truth* should lead us to dismiss that reasoning. Accordingly, I take it that a fully satisfying formal account of the behavior of True and False should be token-sensitive.

As the reader may have noticed, questions λ and λ_1 constitute an (interrogative version of) what Gaifman [18] calls the "two lines puzzle". In fact, Gaifman's *pointer semantics* is an example of a token-sensitive theory of truth which gives up the intersubstitutability property and which yields (in alethic terms) similar conclusions with respect to the status of λ and λ_1 as the intuitive reasoning above. As such, it seems promising to develop a token-sensitive account of True and False on the basis of Gaifman's work. Clearly, doing so is far beyond the scope of this paper.

3.4.4 The Wheeler and Barahona argument

Wheeler and Barahona [56] argued that $HLPE^{3}_{syn}$ cannot be solved in less than three questions. Their argument relies on the following lemma from Information Theory.

[16] Meaning that $T(\overline{\sigma})$ and σ are intersubstitutable (without change of semantic value) in every non opaque context.

(QL) If a question has n possible answers, these answers cannot distinguish $m > n$ different possibilities.

Using QL, we see that when True and False have a three-valued answering repertoire, we cannot solve the four roads riddle by asking a single question. In a similar vein—though in a more complicated setting—Wheeler and Barahona appealed to QL to argue that $HLPE^3_{syn}$ cannot be solved in less than 3 questions, where they assumed that True and False have a three-valued answering repertoire.

When True and False have a four-valued answering repertoire however, QL tells us that we *may* be able to solve the four roads riddle by asking a single question. Of course, whether or not we are actually able to do so depends on our ability to find questions such that their four possible answers are correlated to the four relevant states of affairs in a suitable way. We showed how to solve the four roads riddle by a single question. So, by moving from a three- to a four-valued answering repertoire, we can escape the QL based conclusion that "the four roads riddle cannot be solved in a single question". Similarly, by moving from a three- to a four-valued answering repertoire, we escaped the conclusion of [56] that $HLPE^3_{syn}$ cannot be solved in less than three questions.

Clearly then, the number of questions that is needed to solve "Smullyan like riddles" crucially depends on the number of answers that True and False have available. For instance, QL establishes that the *five roads riddle*—which is defined just as you expect it to be—cannot be solved (in one question) when True and False have a four-valued answering repertoire. However, QL leaves open the possibility that the five roads riddle can be solved in a setting in which True and False are assumed to have a five-valued answering repertoire. Here is such a setting. Assume that True and False are not omniscient and that, besides answering with 'yes', 'no', 'both' and 'neither', they *remain silent* when they are asked a question to which they do not know the answer. So, they now have a five-valued answering repertoire. Here is how to solve the five roads riddle. Let $p_1, \ldots, p_5$ be five sentences such that p_i states that the i^{th} road is good and let $p_?$ be an unknowable (by True and False) sentence. The answer to question π, whose structure—a biconditional flanked by an atomic statement and a statement in disjunctive normal form—mirrors the structure of question γ that was used to solve the four roads riddle, allows you to find out which of the five roads is good:

$$\pi : a = g_T \leftrightarrow (N(a,\pi) \wedge p_1) \vee (Y(a,\pi) \wedge p_2) \vee (p_? \wedge p_3) \vee p_4$$

When p_3 is false, True and False know that $(p_? \wedge p_3)$ is false and so, depending on whether p_1, p_2, p_4 or p_5 is true, question π receives a similar treatment as question γ: the answers 'neither', 'both', 'yes' and 'no' are received just in case, respectively, p_1, p_2, p_4 and p_5 is true. However, when p_3 is true, the answerhood conditions of π reduce to those of $a = g_T \leftrightarrow (p_? \wedge p_3)$. As True and False do not know the truth value of $p_?$, they do not know the truth value of $(p_? \wedge p_3)$ and so they do not know the truth value of $a = g_T \leftrightarrow (p_? \wedge p_3)$. As a consequence, they must remain silent on π. Puzzle solved.

Let me conclude this section by stating a puzzle, $HLPE^{4*}_{syn}$, which I do not know how to solve (in two questions) and which is not unsolvable on the basis of QL. $HLPE^{4*}_{syn}$ is defined just like $HLPE^4_{syn}$, apart from the following

difference. The gods react to your questions with 'huh', 'duh', 'da' and 'ja'. As before, 'da' and 'ja' mean, in some order, 'yes' and 'no'. But now 'huh' and 'duh' mean, in some order, 'neither' and 'both'. Can we solve $HLPE^{4*}_{syn}$ in two questions?

3.5 Concluding remarks

We put forward an alternative conception of how True and False answer yes-no questions, resulting in a four-valued answering repertoire. We then showed how this conception could be invoked to solve $HLPE^{4}_{syn}$ in two questions. Our four-valued (in contrast to a three-valued) answering repertoire allowed us to escape the argument of Wheeler and Barahona [56], which established that $HLPE^{3}_{syn}$ cannot be solved in less than three questions.

The second part of the paper was concerned with formalizations of (the present and previous) solutions to versions of $HLPE$, that were all presented informally. We showed how—by appealing to Strong Kleene fixed point theories of truth and by working in a restricted setting—to give a formal representation of the solutions to $HLPE$. Although in an important sense our formalization "gets the job done", we discussed some desiderata of a formalization of the behavior of True and False that were not met by the one that was presented. To develop a formal account of True and False that meets these desiderata—i.e., a token-sensitive account for an expressive complete language in which True and False can be understood as, respectively, "always speaking truly" and "always speaking falsely"—is postponed to future work.

Chapter 4

Assertoric Semantics and the Computational Power of Self-Referential Truth

4.1 Abstract

There is no consensus as to whether a Liar sentence is meaningful or not. Still, a widespread conviction with respect to Liar sentences (and other *ungrounded* sentences) is that, whether or not they are meaningful, they are *useless*. The philosophical contribution of this paper is to put this conviction into question. Using the framework of *assertoric semantics*, which is a semantic valuation method for languages of self-referential truth that has been developed by the author, we show that certain computational problems, called *query structures*, can be solved more efficiently by an agent who has self-referential resources (amongst which are Liar sentences) than by an agent who has only classical resources; we establish the *computational power of self-referential truth*. The paper concludes with some thoughts on the implications of the established result for deflationary accounts of truth.

4.2 The Useless Liar Conviction

The aim of this paper is to discredit a conviction—widely shared among philosophers and laypersons alike—concerning *the Liar*, i.e., the following sentence:

(**L**) **L** is not true.

The conviction, which I will call the *Useless Liar Conviction* (*ULC*), can be stated as follows:

ULC: The Liar is useless *within* (our) language.

The term 'within' in the formulation of *ULC* is put in italics to distinguish the sense of usefulness alluded to in *ULC* from another sense of usefulness that can

be explicated as follows. The Liar is regarded, by the author and by lots of other philosophers, as useful because it discredits our Naive Theory of truth[1] and thereby forces us to come up with a better theory of truth. But this last sense of 'useful', useful in teaching us something *about* truth (and language), is to be distinguished from *ULC*'s notion of usefulness, which was described as useful *within* our language. The sense in which the Liar is useless according to *ULC* may be illustrated by the following *ULC argument*:

Arguably, the Liar does not express a proposition. Accordingly, the assertion one makes by uttering the Liar, if any, is an infelicitous or useless one. As it seems clear that the Liar cannot be used to realize any other speech act, the Liar is useless within our language.

The *ULC* argument overlooks, I will argue, the fact that the Liar *can* be used to realize another speech act. I hold that by uttering the Liar with interrogative force, as in (4.1), one *does* ask a genuine question.

$$\text{Is } \mathbf{L} \text{ true ?} \tag{4.1}$$

My reason for holding that (4.1) does not fail as a question is that I think that (4.1) has a truthful (correct) answer, which has the following form.

$$\text{(4.1) can neither be answered positively nor negatively} \tag{4.2}$$

At this point a *ULC* proponent may shrug his shoulders. "Fair enough," he may say, "let's grant that your *Liar question* is a (genuine) question; it is—in a sense of usefulness which is clear enough—a useless question anyway". Wittgenstein was, at some point, shrugging his shoulders about the Liar in a way quite similar to the *ULC* proponent.

> If the question is whether this [The Liar] is a statement at all, I reply: You may say that it is not a statement. Or you may say that it *is* a statement, but a useless one. ([68], p209)

In this paper I will discredit the *ULC* by showing that the Liar (question) is useful *within* our language. To be sure, I do not claim that asking the Liar in isolation, i.e., that asking (4.1) is useful. Asking (4.1) is useless because the truthful answer to (4.1) does not convey any information (about the world). Similarly, I regard the question whether snow is white or not white to be a useless one, as no information is conveyed by a truthful answer to that question either. The usefulness of the Liar will be established by showing that *in a language* which contains the Liar (and the *Truthteller*[2]), information can be retrieved more efficiently than in a language without such self-referential resources. Or, to use a nice slogan, the usefulness of the Liar will be established by revealing the *Computational Power of Self-Referential Truth* (*CPSRT*). Before we explain in which sense self-referential truth has computational power, we sketch the contours of the semantic framework, that of *assertoric semantics*, in which the *CPSRT* result is obtained.

[1] The Naive Theory of truth may be thought of as the (smallest) class consisting of all instances of the T-schema and closed under the inference rules of classical logic.

[2] By the Truthteller we mean the sentence: (**T**) **T** is true.

Assertoric semantics is a semantic valuation method for languages of self-referential truth which is developed by the author. Assertoric semantics is a *four valued* semantics and the assertoric value of a sentence σ reports whether or not it is *allowed to assert* σ and also, whether or not it is *allowed to deny* σ. The assertoric norms that are constitutive for assertoric semantics are captured by the *assertoric formula*.

Assertoric formula: it is allowed to assert (deny) σ just in case by asserting (denying) σ one does not assert a falsehood, deny a truth, or contradict oneself.

To save some notation, in the rest of this paper we will use 'it is allowed to assert σ' as shorthand for 'it is allowed to assert σ according to the norms that are expressed by the assertoric formula'. A similar remark applies to the phrase 'it is allowed to deny σ'. In general, the assertoric value of a sentence is determined by two factors: by the world on the one hand and by the *assertoric rules* of one's language on the other. Roughly, the assertoric rules of a language are the inferential rules—for the logical constants, including the truth predicate, of the language—*under an assertoric interpretation*. Let us sketch how the Liar receives its semantic value according to assertoric semantics. If one asserts the Liar, i.e., if one asserts '**L** is not true', one is committed (by the assertoric rules) to the denial of '**L** is true' which commits one to the denial of **L**, i.e., to the denial of the Liar. So, in asserting the Liar one is committed, by the assertoric rules, to the assertion and to the denial of the Liar: we will say that in asserting the Liar, *one contradicts oneself*. Accordingly, it is not allowed (in any world) to assert the Liar. Similarly, it is not allowed to deny the Liar. Thus, it is neither allowed to assert nor to deny the Liar (in any world). A similar argument gives us the assertoric value of the Truthteller: in any world, it is allowed to assert, and it is allowed to deny, the Truthteller. In the actual world, it is allowed to assert 'snow is white' as in doing so one does not violate any of the three constraints mentioned by the assertoric formula. It is not allowed to deny 'snow is white' in the actual world though, as in doing so, one denies a truth. Similarly, it is allowed to deny but not to assert 'snow is black' in the actual world. In the next section, the framework of assertoric semantics will be explicated in more detail. Let us now turn to a sketch of the *CPSRT* result.

The *CPSRT* is nicely illustrated by the following riddle. Suppose that some flag is either red, yellow, green or black and that it is your task to find out which of those four colors the flag has. You can ask yes-no questions to an *oracle*, which is an omniscient entity that *truthfully* answers all and only yes-no questions. Thus, questions like 'what is the color of the flag?' are not answered by the oracle. By a truthful answer to σ we mean an answer that reveals the assertoric value of σ. For instance, (4.2) is a truthful answer to (4.1), as it reveals that it is neither allowed to assert nor to deny the Liar. The riddle is as follows. How many questions do you have to ask the oracle in order to be sure that, after the oracle has answered your questions, you know the flag's color? Two *classical* questions will certainly suffice; first ask whether the flag is (red or yellow) after which you either ask whether the flag is red (if the oracle answered with 'yes') or whether the flag is green (if the oracle answered with 'no'). However, when we have a language which contains self-referential resources, such as Liars and Truthtellers, we can come up with a *single* question that allows us to determine the color of the flag. With **L** and **T** names for the Liar and the Truthteller

respectively, an example of such a "one shot question" is the following question.

Is it the case that: (**L** is not true and the flag is red) or (**T** is true and the flag is yellow) or (the flag is green)?

Thus, though asking the Liar question by itself is not useful, *within a language*, that is in combination with other questions, the Liar can be *used* to create very efficient questions.[3]

This paper is organized as follows. In Section 4.3 we sketch the main ideas of assertoric semantics and we develop a version of assertoric semantics which is tailor made for deriving the *CPSRT* result. Section 4.4 is devoted to the *CPSRT*. There we define the notion of a *query structure*, in which an agent can gain information by querying an oracle. The oracle answers the agent's questions *truthfully*, which is formalized by assuming that the answer of the oracle to a question σ is the assertoric value of σ. The *CPSRT* will be established by showing that the *magical query complexity* of certain query structures is lower than their *classical query complexity*. The magical query complexity is a measure of the efficiency of information retrieval by an agent which has self-referential resources (questions) at his disposal, while the classical query complexity measures the efficiency of information retrieval of an agent without these self-referential resources. In Section 4.5 we step back and reflect on the significance of the *CPSRT* result. In particular, we address the question in what sense the *CPSRT* result is really a result about computation; is our notion of magical query complexity a *natural* notion of computational (query) complexity? Finally, we discuss the implications of the *CPSRT* result for deflationary accounts of truth.

4.3 Assertoric semantics

4.3.1 Quotational closures and truth languages

In this paper, we will develop an assertoric semantics for a (quantifier free) *truth language* $\mathcal{L}$. A truth language $\mathcal{L} = \langle \bar{L}, \pi \rangle$ consists of a (quantifier free) quotational closure $\bar{L}$ together with a *reference list* π. These notions are defined as follows.

Definition 4.1 Quotational closures and their classical fragment
A *quotational closure* $\bar{L}$ has $\{\vee, \wedge, \neg, T\}$ as its logical symbolism, where 'T' is a unary truth predicate. Its non-logical symbolism consists, amongst others, of:

1. A set $P = \{p_1, p_2, \ldots\}$ of propositional atoms.
2. A finite set $C = \{c_1, c_2, \ldots, c_n\}$ of *non-quotational constant symbols*.

Besides the non-quotational constant symbols C, $\bar{L}$ also contains a set of *quotational constant symbols* $[C] = Con(\bar{L}) - C$. The set $Con(\bar{L})$, consisting of all

[3]Inspiration for this paper is taken from Rabern and Rabern [43], who solve (in natural language) a 1-out-of-3 riddle by asking a single question to an oracle. For an evaluation of their solution in a completely distinct formal framework than that of this paper, see Wintein [58].

the constant symbols (quotational or not) of $\bar{L}$ is jointly defined with $Sen(\bar{L})$[4], the set of sentences of $\bar{L}$. $Sen(\bar{L})$ and $Con(\bar{L})$ are the smallest sets satisfying:

- $P \subseteq Sen(\bar{L})$, $C \subseteq Con(\bar{L})$.
- $\alpha \in Sen(\bar{L}) \Rightarrow [\alpha] \in Con(\bar{L})$.
- $t \in Con(\bar{L}) \Rightarrow T(t) \in Sen(\bar{L})$.
- $\alpha, \beta \in Sen(\bar{L}) \Rightarrow \neg\alpha, (\alpha \wedge \beta), (\alpha \vee \beta) \in Sen(\bar{L})$.

We let $L_P = \langle P, \{\vee, \wedge, \neg\}\rangle$ denote the propositional language over P. L_P is called the *classical fragment* of $\bar{L}$. □

A quotational closure thus has, corresponding with each $\sigma \in Sen(\bar{L})$, a quotational constant symbol $[\sigma]$. In a truth language $\mathcal{L} = \langle \bar{L}, \pi\rangle$, we have that $\pi([\sigma]) = \sigma$ and we say that the quotational constant symbol $[\sigma]$ *refers to* σ.

Definition 4.2 Truth languages, their classical fragment and worlds
A *truth language* $\mathcal{L} = \langle \bar{L}, \pi\rangle$ is a pair consisting of a quotational closure $\bar{L}$ and a function $\pi : Con(\bar{L}) \to Sen(\bar{L})$, called a *reference list*, which satisfies

$$\pi([\sigma]) = \sigma \qquad \text{for all } \sigma \in Sen(\bar{L})$$

When $\mathcal{L}$ is a truth language, we let $Con(\mathcal{L}) = Con(\bar{L})$ and we let $Sen(\mathcal{L}) = Sen(\bar{L})$. Also, we say that L_P *is the classical fragment of* $\mathcal{L}$. By a *world* $w \in W = \mathcal{P}(P)$ we mean a set of propositional atoms. We think of a world as a *model* for L_P. □

Throughout the paper, 'λ' and 'τ' will be used as non-quotational constant symbols satisfying (4.3).

$$\pi(\lambda) = \neg T(\lambda), \qquad \pi(\tau) = T(\tau) \tag{4.3}$$

Under this convention, $\neg T(\lambda)$ models the Liar, whereas $T(\tau)$ models the Truthteller.

4.3.2 Assertoric values, -rules and -trees

In this paper we will only be concerned with assertoric semantics for truth languages as defined in Section 4.3.1. In what follows, $\mathcal{L}$ will denote an arbitrary such language.

An *assertoric valuation function* is a function $\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$, taking a sentence σ and a world w as input and returning an element of $\{0,1\}^2$, which is interpreted as the *assertoric value* of σ in w. That is, with $\mathcal{V}(\sigma, w) = (x, y)$, we have that:

$$x = 1 \Leftrightarrow \text{the assertion of } \sigma \text{ is allowed in } w\ .$$
$$y = 1 \Leftrightarrow \text{the denial of } \sigma \text{ is allowed in } w.$$

Here and elsewhere in the text 'allowed' is used as shorthand for 'allowed according to the assertoric norms which are expressed by the assertoric formula'.

[4]Sentences that are constructed using connectives in $\{\rightarrow, \leftrightarrow\}$ are officially in the metalanguage; the translation of them to "official sentences" is achieved in the usual manner.

Whether or not is is allowed to assert (deny) a sentence σ in a world w is determined by two factors; by the world w on the one hand and by the *assertoric rules* of $\mathcal{L}$ on the other. The assertoric rules of $\mathcal{L}$ are basically the inferential rules of $\mathcal{L}$ *under an assertoric reading*. In order to present the rules and their interpretation, we define $\mathcal{X}$, the set of *assertoric sentences of* $\mathcal{L}$.

$$\mathcal{X} = \{X_\sigma \mid X \in \{A, D\}, \sigma \in Sen(\mathcal{L})\}$$

A_σ symbolizes the *assertion* of σ, while D_σ symbolizes the *denial* of σ. We can rephrase the interpretation of an assertoric valuation function as follows. With $\mathcal{V}(\sigma, w) = (x, y)$, we have:

$x = 1 \Leftrightarrow A_\sigma$ is allowed in w.
$y = 1 \Leftrightarrow D_\sigma$ is allowed in w.

A sentence $\sigma \in Sen(\mathcal{L})$ has either one of the following five forms: $p, (\alpha \wedge \beta), (\alpha \vee \beta), \neg\alpha$ or $T(t)$. The form of an assertoric sentence X_σ is specified by its sign, $X \in \{A, D\}$, and by the form of σ. Thus, there are 10 possible forms of assertoric sentences or, in other words, there are ten *assertoric forms*. With each of the assertoric forms we associate an *assertoric rule*. An assertoric rule is either of conjunctive type $\sqcap$ or of disjunctive type $\sqcup$. Depending on its type, an assertoric rule is depicted in either one of the following two ways.

$$\frac{X_\sigma}{\Pi(X_\sigma)}\sqcup \qquad \frac{X_\sigma}{\Pi(X_\sigma)}\sqcap$$

Formally, the assertoric rule $\mathcal{R}(\mathsf{AF})$ associated with assertoric form AF can be thought of as a rule associating each assertoric sentence of form AF with its set of *immediate* π *sentences* $\Pi(X_\sigma)$, in either a conjunctive way ($\sqcap$) or in a *disjunctive way* ($\sqcup$). The ten assertoric rules for $\mathcal{L}$ are, together with their type, displayed in the following table.

$\frac{A_{\neg\alpha}}{\{D_\alpha\}}\sqcap$	$\frac{D_{\neg\alpha}}{\{A_\alpha\}}\sqcap$
$\frac{A_{(\alpha\vee\beta)}}{\{A_\alpha, A_\beta\}}\sqcup$	$\frac{D_{(\alpha\vee\beta)}}{\{D_\alpha, D_\beta\}}\sqcap$
$\frac{A_{(\alpha\wedge\beta)}}{\{A_\alpha, A_\beta\}}\sqcap$	$\frac{D_{(\alpha\wedge\beta)}}{\{D_\alpha, D_\beta\}}\sqcup$
$\frac{A_{T(t)}}{\{A_{\pi(t)}\}}\sqcap$	$\frac{D_{T(t)}}{\{D_{\pi(t)}\}}\sqcap$
$\frac{A_p}{\{A_p\}}\sqcap$	$\frac{D_p}{\{D_p\}}\sqcap$

The assertoric rules are thus the usual tableau rules for $\wedge, \vee, \neg$ in terms of signed statements (as in Smullyan[50]), augmented with rules that govern the truth predicate and the (trivial) rules for propositional atoms. It will be convenient to apply the notion of type ($\sqcup$ or $\sqcap$) not only to rules but also to assertoric sentences in accordance with the table above. For instance, we will say that the type of a sentences of form $A_{(\alpha\vee\beta)}$ and $D_{(\alpha\wedge\beta)}$ is $\sqcup$. We distinguish three readings of the assertoric rules. We call them the $\mathcal{L}(\Rightarrow)$, $\mathcal{L}(\Leftarrow)$ and $\mathcal{L}(\Leftrightarrow)$ rules respectively.

Definition 4.3 The $\mathcal{L}(\Rightarrow)$, $\mathcal{L}(\Leftarrow)$ and $\mathcal{L}(\Leftrightarrow)$ rules
Let AF be an assertoric form and let $\mathcal{R}(\mathsf{AF})$ be the associated assertoric rule. Depending on the type, $\sqcup$ or $\sqcap$, of $\mathcal{R}(\mathsf{AF})$, *the* $\mathcal{L}(\Rightarrow)$ *rule* associated with $\mathcal{R}(\mathsf{AF})$ is as follows. For any X_σ of form AF:

$\sqcap$: one is committed to $X_\sigma \Rightarrow$ one is committed to Y_α for all $Y_\alpha \in \Pi(X_\sigma)$.

$\sqcup$: one is committed to $X_\sigma \Rightarrow$ one is committed to Y_α for some $Y_\alpha \in \Pi(X_\sigma)$.

The $\mathcal{L}(\Leftarrow)$ rule and $\mathcal{L}(\Leftrightarrow)$ rule associated with $\mathcal{R}(\mathsf{AF})$ are obtained by replacing, in the $\mathcal{L}(\Rightarrow)$ rule, '$\Rightarrow$' with '$\Leftarrow$' and '$\Leftrightarrow$' respectively. Each of the ten assertoric forms has an $\mathcal{L}(\Rightarrow)$ rule associated with it and by *the* $\mathcal{L}(\Rightarrow)$ *rules* we mean the collection of ten $\mathcal{L}(\Rightarrow)$ rules. The notions of $\mathcal{L}(\Leftarrow)$ *rules* and $\mathcal{L}(\Leftrightarrow)$ *rules* are defined similarly. The term $\mathcal{L}(\cdot)$ *rules* will be used as a variable that ranges over the $\mathcal{L}(\Rightarrow)$, $\mathcal{L}(\Leftarrow)$ and $\mathcal{L}(\Leftrightarrow)$ rules. □

As an example, consider the rule associated with $A_{\neg\alpha}$. As $A_{\neg\alpha}$ has type $\sqcap$ and as $\Pi(A_{\neg\alpha}) = \{D_\alpha\}$, the $\mathcal{L}(\Rightarrow)$ rule associated with the assertion of a negation reads as follows:

$$\text{one is committed to } A_{\neg_\alpha} \Rightarrow \text{one is committed to } D_\alpha.$$

So, if one is committed to the assertion of $\neg\alpha$, one is committed to the denial of α. There are two ways in which one can be committed to the assertion (denial) of a sentence σ. One is by actually asserting (denying) σ. The other is by asserting or denying some other sentence, say σ_0, which then induces a commitment to the assertion (denial) of σ via the assertoric rules. For instance, if I assert $(\alpha \wedge \beta)$ I am committed to the assertion of $(\alpha \wedge \beta)$, as I (actually) asserted $(\alpha \wedge \beta)$. However, I am also committed, via the assertoric rules, to the assertion of α and to the assertion of β, although I did not actually assert α or β. Before we expand on these remarks, we define the notion of a valuation function $\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$ *validating* a set of $\mathcal{L}(\cdot)$ rules. In this definition, we use the two *projection functions* $\mathcal{V}_X : Sen(\mathcal{L}) \times W \to \{0,1\}$ of $\mathcal{V}$ where, with $\mathcal{V}(\sigma, w) = (x, y)$ we have that $\mathcal{V}_A(\sigma, w) = x$ and $\mathcal{V}_D(\sigma, w) = y$.

Definition 4.4 Validity of $\mathcal{L}(\cdot)$ rules with respect to $\mathcal{V}$
Let $\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$. We say that the $\mathcal{L}(\Rightarrow)$ *rules are valid with respect to* $\mathcal{V}$ just in case we have, for each assertoric sentence $X_\sigma \in \mathcal{X}$ and for each world $w \in W$ that:

$$\text{type of } X_\sigma \text{ is } \sqcap\text{: } \mathcal{V}_X(\sigma, w) = 1 \Rightarrow \mathcal{V}_Y(\alpha, w) = 1 \text{ for all } Y_\alpha \in \Pi(X_\sigma). \tag{4.4}$$

$$\text{type of } X_\sigma \text{ is } \sqcup\text{: } \mathcal{V}_X(\sigma, w) = 1 \Rightarrow \mathcal{V}_Y(\alpha, w) = 1 \text{ for some } Y_\alpha \in \Pi(X_\sigma). \tag{4.5}$$

The conditions for the validity of the $\mathcal{L}(\Leftarrow)$ and $\mathcal{L}(\Leftrightarrow)$ rules with respect to $\mathcal{V}$ are obtained by replacing, in (4.4) and (4.5), '$\Rightarrow$' with, respectively, '$\Leftarrow$' and '$\Leftrightarrow$'. □

The $\mathcal{L}(\Rightarrow)$ rules transmit assertoric commitments in the sense that the commitment to the assertion or denial of a certain sentence is translated, via the $\mathcal{L}(\Rightarrow)$ rules, into assertoric commitments with respect to other sentences. In

order to keep track of the sum total of assertoric commitments that are involved with the assertion or denial of a certain sentence σ, we define, for each $\sigma \in Sen(\mathcal{L})$, its two *assertoric trees*, consisting of σ's *assertion tree* $\mathfrak{T}^{\sigma}_{A}$ and its *denial tree* $\mathfrak{T}^{\sigma}_{D}$. We think of an assertoric tree $\mathfrak{T}^{\sigma}_{X}$ as the set which consists of all the *branches of* X_σ.

Definition 4.5 Branches of X_σ
A set $B \subseteq \mathcal{X}$ is a *branch* of X_σ just in case conditions 1, 2 , 3 and 4 hold.

1. $X_\sigma \in B$
2. ($Y_\alpha \in B$ and Y_α has type $\sqcap$) $\Rightarrow$ $Z_\beta \in B$ for all $Z_\beta \in \Pi(Y_\alpha)$
3. ($Y_\alpha \in B$ and Y_α has type $\sqcup$) $\Rightarrow$ $Z_\beta \in B$ for some $Z_\beta \in \Pi(Y_\alpha)$
4. For no $S \subset B$, condition 1,2 and 3 are satisfied.

A branch of X_σ is thus a set containing X_σ which is (downwards) saturated under the $\mathcal{L}(\Rightarrow)$ rules and minimal in the sense that no strict subset of B contains X_σ and is (downwards) saturated under the $\mathcal{L}(\Rightarrow)$ rules. □

We are now ready to define the two assertoric trees, $\mathfrak{T}^{\sigma}_{A}$ and $\mathfrak{T}^{\sigma}_{D}$, of a sentence $\sigma \in Sen(\mathcal{L})$.

Definition 4.6 Assertoric trees
Let $\sigma \in Sen(\mathcal{L})$. Its *assertion tree* $\mathfrak{T}^{\sigma}_{A}$ and its *denial tree* $\mathfrak{T}^{\sigma}_{D}$ are defined as follows:

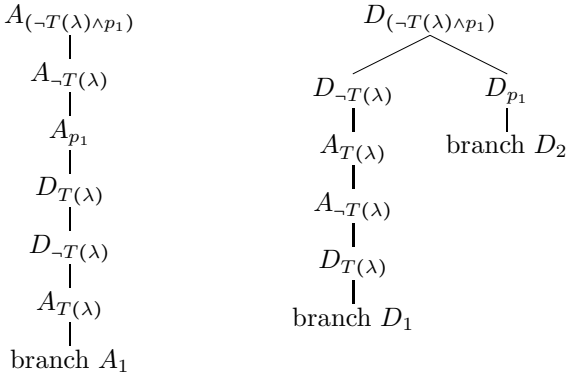
$$\mathfrak{T}^{\sigma}_{A} = \{B \mid B \text{ is a branch of } A_\sigma\}$$

$$\mathfrak{T}^{\sigma}_{D} = \{B \mid B \text{ is a branch of } D_\sigma\}$$

An assertoric tree is thus a set of sets of assertoric sentences. □

Although assertoric trees are officially sets of sets of assertoric sentences, we will depict them as "genuine trees". For instance, we do so in the following example.

Example 4.1 Assertoric trees
Below we depict $\mathfrak{T}^{\gamma}_{A}$ and $\mathfrak{T}^{\gamma}_{D}$, where $\gamma := (\neg T(\lambda) \wedge p_1) \in Sen(\mathcal{L})$.



Due to the finiteness of C, an assertoric tree is a *finite* object, being a finite set whose elements are finite sets of sentences. □

The remark made at the end of Example 4.1 is Proposition 4.1.

Proposition 4.1 $\mathfrak{T}^{\sigma}_{X}$ is a finite set whose elements are finite sets.
Proof: See appendix. Roughly, the proof follows from the assumption that C is finite and that $\mathcal{L}$ is quantifier free. □

4.3.3 Inducing valuation functions by closure conditions

In this subsection, we describe how *closure conditions for branches* are used to induce a valuation function $\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$. Here, we describe the general form of these inducements, while the next subsection describes a concrete such inducement.

By *closure conditions* for branches, we mean necessary and sufficient conditions which specify under which circumstances a branch is called *closed in a world* w, while a branch is called *open in* w just in case it is not closed in w. We write $C_w(B)$ and $O_w(B)$ as shorthand for 'branch B is closed in world w' and 'branch B is open in world w' respectively. *Closure conditions for branches define closure conditions for assertoric trees* according to the following schema.

$$C_w(\mathfrak{T}^{\sigma}_{X}) \Leftrightarrow_{def} C_w(B) \text{ for all } B \in \mathfrak{T}^{\sigma}_{X} \tag{4.6}$$

Equivalently, we can phrase this as:

$$O_w(\mathfrak{T}^{\sigma}_{X}) \Leftrightarrow_{def} O_w(B) \text{ for some } B \in \mathfrak{T}^{\sigma}_{X} \tag{4.7}$$

Closure conditions induce a valuation function $\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$ according to the following definition.

Definition 4.7 $\mathcal{V}$ induced by closure conditions
Closure conditions for branches induce a valuation function $\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$. We define $\mathcal{V}$ in terms of its projector functions $\mathcal{V}_X$ as follows:

$$\mathcal{V}_X(\sigma, w) = 1 \Leftrightarrow O_w(\mathfrak{T}^{\sigma}_{X}),$$

where the closure conditions for trees derive from the closure conditions for branches as specified by equation (4.6) or (4.7). □

Any set of closure conditions can be used to induce a valuation function in accordance with Definition 4.7. However, not any set of closure conditions induces a valuation function that can properly be called an *assertoric* valuation function. In order to induce an assertoric valuation function, the closure conditions should be an adequate formal representation of the assertoric norms expressed by the assertoric formula. In other words, the closure conditions should *capture the assertoric formula*. Closure conditions which capture the assertoric formula allow us to explain the closure of an assertoric tree along the following lines. In asserting σ one takes up assertoric commitments that are summarized by the assertion tree of σ. The closure of the assertion tree of σ means that one is not able, judged by the standards of the assertoric formula, to live up to the involved assertoric commitments. For instance, if the assertion tree of σ is closed in world w it is the case that, intuitively, in asserting σ in w one is committed to the assertion of a falsehood, the denial of a truth, or to contradict oneself. Similar remarks apply to the (closure of the) denial tree of σ. Conversely, if

the assertion (denial) tree of σ is open, one can live up to the commitments of asserting (denying) σ and, accordingly, it is allowed to assert (deny) σ.

In the next section we will define the *assertoric closure conditions*. The assertoric closure conditions capture the assertoric formula and are used to induce the valuation function $\mathcal{V}^{\mathsf{as}}$. As we will see, $\mathcal{V}^{\mathsf{as}}$ validates the $\mathcal{L}(\Rightarrow)$ rules, but not the $\mathcal{L}(\Leftarrow)$ rules.

4.3.4 The assertoric valuation function $\mathcal{V}^{\mathsf{as}}$

We now define the assertoric closure conditions for sets of assertoric sentences. By extension, we thereby define closure conditions for branches.

Definition 4.8 Assertoric closure conditions
Let $w \in W$ and let $S \subseteq \mathcal{X}$. S is *open in w according to the assertoric closure conditions*, denoted $O^{\mathsf{as}}_w(S)$, iff S is not *closed in w* according to those conditions. S is *closed in w according to the assertoric closure conditions*, denoted $C^{\mathsf{as}}_w(S)$, iff the disjunction of 1, 2 and 3 is true.

1. There is a $p \in P$ such that $A_p \in S$ and $p \notin w$.
2. There is a $p \in P$ such that $D_p \in S$ and $p \in w$.
3. There is a $\sigma \in Sen(\mathcal{L})$ such that $A_\sigma \in S$ and $D_\sigma \in S$. □

The assertoric closure conditions capture the assertoric formula which says that one is allowed to assert (deny) a sentence σ just in case in asserting σ one does not assert a falsehood (condition 1), deny a truth (condition 2) or contradict oneself (condition 3). We let $\mathcal{V}^{\mathsf{as}}$ be the valuation function which is induced by the assertoric closure conditions according to Definition 4.7.

Proposition 4.2 $\mathcal{V}^{\mathsf{as}}$ validates the $\mathcal{L}(\Rightarrow)$ rules
Proof: We illustrate that the $\mathcal{L}(\Rightarrow)$ rule of $A_{\alpha\wedge\beta}$ is valid with respect to $\mathcal{V}^{\mathsf{as}}$. The $\mathcal{L}(\Rightarrow)$ validity of the other assertoric rules can be shown by a similar argument. Thus, we will prove that:

$$\mathcal{V}^{\mathsf{as}}_A(\alpha \wedge \beta, w) = 1 \Rightarrow \mathcal{V}^{\mathsf{as}}_A(\alpha, w) = 1 \text{ and } \mathcal{V}^{\mathsf{as}}_A(\beta, w) = 1$$

We proceed by reductio. Suppose that $\mathcal{V}^{\mathsf{as}}_A(\alpha\wedge\beta, w) = 1$, i.e. that $O^{\mathsf{as}}_w(\mathfrak{T}^{\alpha\wedge\beta}_A)$ and suppose that not $\mathcal{V}^{\mathsf{as}}_A(\alpha, w) = 1$, i.e. that $C^{\mathsf{as}}_w(\mathfrak{T}^{\alpha}_A)$. Observe that for all $B \in \mathfrak{T}^{\alpha\wedge\beta}_A$ there exists a $B' \in \mathfrak{T}^{\alpha}_A$ such that $B' \subset B$. From $O^{\mathsf{as}}_w(\mathfrak{T}^{\alpha\wedge\beta}_A)$ it follows that there exists a branch, say B, of $\mathfrak{T}^{\alpha\wedge\beta}_A$ such that $O^{\mathsf{as}}_w(B)$. By our observation, B is the superset of some branch B' of $\mathfrak{T}^{\alpha}_A$. As, by hypothesis, we have that $C^{\mathsf{as}}_w(\mathfrak{T}^{\alpha}_A)$ it follows that $C^{\mathsf{as}}_w(B')$. As it is impossible to have $B' \subset B$ with $C^{\mathsf{as}}_w(B')$ and $O^{\mathsf{as}}_w(B)$ we are done. □

Example 4.2 $\mathcal{V}^{\mathsf{as}}$ does not validate the $\mathcal{L}(\Leftarrow)$ rules
Let $\pi(\tau) = T(\tau)$, i.e. $T(\tau)$ is a *Truthteller*. Observe that, with w an arbitrary world, we have that:

$$\mathcal{V}^{\mathsf{as}}(T(\tau), w) = (1,1), \qquad \mathcal{V}^{\mathsf{as}}(\neg T(\tau), w) = (1,1)$$
$$\mathcal{V}^{\mathsf{as}}(T(\tau) \wedge \neg T(\tau), w) = (0,1)$$

These assertoric values falsify the $\mathcal{L}(\Leftarrow)$ rule of $A_\wedge$, as:

$$\mathcal{V}^{\mathsf{as}}_A(\alpha \wedge \beta, w) = 1 \not\Leftarrow (\mathcal{V}^{\mathsf{as}}_A(\alpha, w) = 1 \text{ and } \mathcal{V}^{\mathsf{as}}_A(\beta, w) = 1)$$

Thus, $\mathcal{V}^{\mathsf{as}}$ does not validate the $\mathcal{L}(\Leftarrow)$ rules and so, *a fortiori* neither does $\mathcal{V}^{\mathsf{as}}$ validate the $\mathcal{L}(\Leftrightarrow)$ rules. □

Interestingly, an (arbitrary) valuation function $\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$ validates the $\mathcal{L}(\Leftrightarrow)$ rules just in case $\mathcal{V}$ is *compositional* with respect to $FOUR_\leq$, where the structure $FOUR_\leq = \langle \{0,1\}^2, \leq \rangle$, familiar from o.a. ([7]), partially orders the four assertoric values according to the following Hasse diagram.

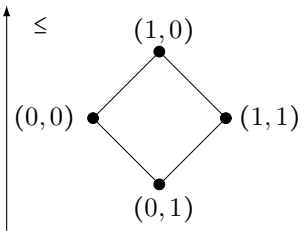


Figure 4.1: $Four_\leq$

Definition 4.9 $FOUR_\leq$ **compositionality**
$\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$ is said to be $FOUR_\leq$ *compositional* just in case conditions *a*, *b*, *c* and *d* are fulfilled, where $t \in Con(\mathcal{L})$ and $w \in W$.

a. 1. $\mathcal{V}(\alpha, w) = (1,0) \Leftrightarrow \mathcal{V}(\neg\alpha, w) = (0,1)$
2. $\mathcal{V}(\alpha, w) = (0,1) \Leftrightarrow \mathcal{V}(\neg\alpha, w) = (1,0)$
3. $\mathcal{V}(\alpha, w) = (0,0) \Leftrightarrow \mathcal{V}(\neg\alpha, w) = (0,0)$
4. $\mathcal{V}(\alpha, w) = (1,1) \Leftrightarrow \mathcal{V}(\neg\alpha, w) = (1,1)$

b. $\mathcal{V}(T(t), w) = \mathcal{V}(\pi(t), w)$

c. $\mathcal{V}(\alpha \vee \beta, w) = sup_\leq(\{\mathcal{V}(\alpha, w), \mathcal{V}(\beta, w)\})$

d. $\mathcal{V}(\alpha \wedge \beta, w) = inf_\leq(\{\mathcal{V}(\alpha, w), \mathcal{V}(\beta, w)\})$ □

The following proposition is an immediate consequence of Definition 4.4 and Definition 4.9.

Proposition 4.3 $\mathcal{V} : Sen(\mathcal{L}) \times W \to \{0,1\}^2$ **validates the** $\mathcal{L}(\Leftrightarrow)$ **rules just in case** $\mathcal{V}$ **is** $FOUR_\leq$ **compositional**.

Proof: We illustrate the proof by considering conjunction only as all other cases are treated similarly. The $\mathcal{L}(\Leftrightarrow)$ rules of asserting and denying a conjunction are validated by $\mathcal{V}$ just in case we have that, for each world w:

$$\mathcal{V}_A(\alpha \wedge \beta, w) = 1 \Leftrightarrow V_A(\alpha, w) = 1 \text{ and } V_A(\beta, w) = 1 \tag{4.8}$$

$$\mathcal{V}_D(\alpha \wedge \beta, w) = 1 \Leftrightarrow V_D(\alpha, w) = 1 \text{ or } V_D(\beta, w) = 1 \tag{4.9}$$

A row of the following "assertoric table" is understood as a possible way to satisfy equations (4.8) and (4.9) simultaneously. The table as a whole represents all possible ways of doing so.

$\mathcal{V}(\alpha, w)$	$\mathcal{V}(\beta, w)$	$\mathcal{V}(\alpha \wedge \beta, w)$
(0, 0)	(0, 0)	(0, 0)
(0, 0)	(0, 1)	(0, 1)
(0, 0)	(1, 0)	(0, 0)
(0, 0)	(1, 1)	(0, 1)
(0, 1)	(0, 0)	(0, 1)
(0, 1)	(0, 1)	(0, 1)
(0, 1)	(1, 0)	(0, 1)
(0, 1)	(1, 1)	(0, 1)
(1, 0)	(0, 0)	(0, 0)
(1, 0)	(0, 1)	(0, 1)
(1, 0)	(1, 0)	(1, 0)
(1, 0)	(1, 1)	(1, 1)
(1, 1)	(0, 0)	(0, 1)
(1, 1)	(0, 1)	(0, 1)
(1, 1)	(1, 0)	(1, 1)
(1, 1)	(1, 1)	(1, 1)

The values in the table testify that $\mathcal{V}$ is $FOUR_{\leq}$ compositional with respect to $\wedge$. To see that $\mathcal{V}$ is $FOUR_{\leq}$ compositional with respect to the other connectives one reasons similarly. □

And so, Proposition 4.3 and Example 4.2 tell us that $\mathcal{V}^{\mathsf{as}}$ is not $FOUR_{\leq}$ compositional, which can also be observed by inspecting the following assertoric values:

$$\mathcal{V}^{\mathsf{as}}(T(\tau), w) = (1,1), \qquad \mathcal{V}^{\mathsf{as}}(\neg T(\tau), w) = (1,1)$$
$$\mathcal{V}^{\mathsf{as}}(T(\tau) \wedge T(\tau), w) = (1,1), \qquad \mathcal{V}^{\mathsf{as}}(T(\tau) \wedge \neg T(\tau), w) = (0,1)$$

The specified assertoric values illustrate that the $\mathcal{V}^{\mathsf{as}}$ value of a conjunction is not determined by the $\mathcal{V}^{\mathsf{as}}$ values of its conjuncts, which means that $\mathcal{V}^{\mathsf{as}}$ is not ($FOUR_{\leq}$) compositional. The non compositionality of $\mathcal{V}^{\mathsf{as}}$ implies that $\mathcal{V}^{\mathsf{as}}$ cannot be characterized as a Kripkean fixed point (described in [33]) that is obtained by the Strong (or Weak) Kleene scheme, for the Kleene valuation schemes are compositional. The specified assertoric values leave open the possibility to characterize $\mathcal{V}^{\mathsf{as}}$ as Kripkean fixed point that is obtained by the Supervaluation scheme. For the Supervaluation scheme is a non compositional valuation scheme in which the conjunction of a sentence with its negation is always valuated as "false"—where (roughly) "false" is the analogue of our value (0, 1). However, although $\mathcal{V}^{\mathsf{as}}$ is not compositional, it does not have the property of assigning the value $(0,1)$ to each conjunction of a sentence with its negation, as is testified by the Liar sentence $T(\lambda)$:

$$\mathcal{V}^{\mathsf{as}}(T(\lambda) \wedge \neg T(\lambda), w) = (0,0)$$

And so, $\mathcal{V}^{\mathsf{as}}$ cannot be characterized as a Kripkean fixed point that is obtained by the Supervaluation scheme either. For a detailed comparison of Kripkean fixed

point semantics with assertoric semantics, the reader is referred[5] to Wintein [63].

We conclude our discussion of $\mathcal{V}^{\mathsf{as}}$ by observing that, though $\mathcal{V}^{\mathsf{as}}$ cannot be characterized as a Kripkean fixed point, it does have the most interesting property possessed by the Kripkean fixed points: according to $\mathcal{V}^{\mathsf{as}}$ the assertoric value of a sentence σ is equal to the assertoric value of a sentence which says of σ that it is true. That is, $\mathcal{V}^{\mathsf{as}}$ shares the *semantic intersubstitutability of truth* property with the Kripkean fixed points.

Proposition 4.4 $\forall t \in Con(\mathcal{L}), \forall w \in W : \mathcal{V}^{\mathsf{as}}(T(t), w) = \mathcal{V}^{\mathsf{as}}(\pi(t), w)$
Proof:. First observe that the claim is equivalent to the claim that, with $t \in Con(\mathcal{L})$, $w \in W$ and $X \in \{A, D\}$ we have that:

$$O_w^{\mathsf{as}}(\mathfrak{T}_X^{T(t)}) \Leftrightarrow O_w^{\mathsf{as}}(\mathfrak{T}_X^{\pi(t)})$$

$\Rightarrow$ By reductio. Suppose that $O_w^{\mathsf{as}}(\mathfrak{T}_X^{T(t)})$ and suppose that $C_w^{\mathsf{as}}(\mathfrak{T}_X^{\pi(t)})$. Observe that for all $B \in \mathfrak{T}_X^{T(t)}$ there exists a $B' \in \mathfrak{T}_X^{\pi(t)}$ such that $B' \subset B$. From $O_w^{\mathsf{as}}(\mathfrak{T}_X^{T(t)})$ it follows that there exists a branch, say B, of $\mathfrak{T}_X^{T(t)}$ such that $O_w^{\mathsf{as}}(B)$. By our observation, B is the superset of some branch B' of $\mathfrak{T}_X^{\pi(t)}$. As, by hypothesis, we have that $C_w^{\mathsf{as}}(\mathfrak{T}_X^{\pi(t)})$ it follows that $C_w^{\mathsf{as}}(B')$. As it is impossible to have $B' \subset B$ with $C_w^{\mathsf{as}}(B')$ and $O_w^{\mathsf{as}}(B)$ we are done.

$\Leftarrow$. We only give the proof for $X = A$ as the proof for $X = D$ is similar. Proof by reductio. Suppose that $O_w^{\mathsf{as}}(\mathfrak{T}_A^{\pi(t)})$ and that $C_w^{\mathsf{as}}(\mathfrak{T}_A^{T(t)})$. Note that:

$$\mathfrak{T}_A^{T(t)} = \{B \cup \{A_{T(t)}\} \mid B \in \mathfrak{T}_A^{\pi(t)}\}$$

From our reduction hypothesis it thus follows that there has to exist a B such that $O_w^{\mathsf{as}}(B)$ and such that $C_w^{\mathsf{as}}(B \cup \{A_{T(t)}\})$. As $T(t) \notin P$ and as B is (assertorically) open it follows that the closure of $B \cup \{A_{T(t)}\}$ has to be explained by the occurrence of $D_{T(t)}$ in B. As B is saturated under the $\mathcal{L}(\Rightarrow)$ rules, it follows that B contains $D_{\pi(t)}$ as well. And so, as $A_{\pi(t)}$ is the origin of $\mathfrak{T}_A^{\pi(t)}$, B contains both $A_{\pi(t)}$ and $D_{\pi(t)}$ and we have that $C_w^{\mathsf{as}}(B)$. Contradiction. □

In the next section we show how the framework of assertoric semantics, using the valuation function $\mathcal{V}^{\mathsf{as}}$, can be invoked to establish the *Computational Power of Self-Referential Truth*.

4.4 The Computational Power of Self-Referential Truth

4.4.1 Query structures, -strategies and -complexity

In a *query structure*, an agent wants to know the (classical) truth value (true or false) of all sentences in a set $\mathcal{T} \subseteq Sen(L_P)$. Or, which is to say the same,

[5]It turns out, as shown in [63], that $\mathcal{V}^{\mathsf{as}}$ is equivalent to the function that Kripke [33] (implicitly) defined by quantifying over all fixed points. According to this function, the Liar is *paradoxical* as there is no 3-valued (Strong Kleene) fixed point in which it valuates as 1 and also, no fixed point in which it valuates as 0. I did not realize this connection with Kripke's work at the time of writing the present paper.

the agent wants to *decide* his *target knowledge* $\mathcal{T}$. In order to decide $\mathcal{T}$, the agent may invoke his background knowledge $\mathcal{B} \subseteq Sen(L_P)$ in combination with additional information that he can gain by *querying* an oracle. The oracle is an *omniscient* and *truthful entity*. Being omniscient, the oracle knows all facts of the (actual) world, and so he knows the truth value of each sentence of L_P. Being truthful, the oracle, when addressed a *question* $\sigma \in Sen(\mathcal{L})$, truthfully reveals the assertoric possibilities with respect to σ. Thus, the answer of the oracle to σ is equal to $\mathcal{V}^{\mathsf{as}}(\sigma, w_@)$, where $w_@$ is the actual world. Before we proceed, let us formally define the notion of a query structure.

Definition 4.10 Query structure
We say that a pair $\langle \mathcal{B}, \mathcal{T} \rangle$, consisting of $\mathcal{B} \subseteq Sen(L_P)$ and $\mathcal{T} \subseteq Sen(L_P)$, is a *query structure* just in case:

1. $\mathcal{B} \cup \mathcal{T}$ is consistent. $|\mathcal{T}|$ is finite.
2. $\exists \sigma \in \mathcal{T} : \mathcal{B} \nvdash_{L_P} \sigma$ and $\mathcal{B} \nvdash_{L_P} \neg\sigma$. (non-triviality)

Where a set $S \subseteq Sen(L_P)$ is consistent just in case there is a $\sigma \in Sen(L_P)$ such that $S \nvdash_{L_P} \sigma$. Here, $S \vdash_{L_P} \sigma$ means that $\sigma \in Sen(L_P)$ follows from $S \subseteq Sen(L_P)$ by propositional logic. □

We distinguish between two types of agents. A *classical agent* $\mathbf{A_{cla}}$ speaks L_P (and only L_P), while a *magical agent* $\mathbf{A_{ma}}$ speaks some[6] truth language $\mathcal{L}$. More concretely, $\mathbf{A_{cla}}$ can query the oracle by asking him questions in $Sen(L_P)$ while $\mathbf{A_{ma}}$ can ask the oracle questions in $Sen(\mathcal{L})$.

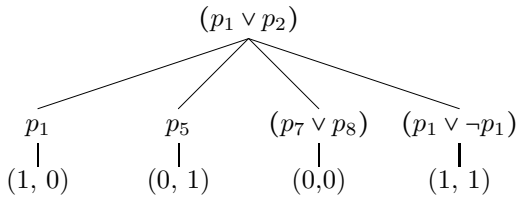
An *n-query strategy* $\mathcal{S}$ of an arbitrary agent $\mathbf{A}$ is a plan of $\mathbf{A}$ to ask n consecutive questions to the oracle, where the m^{th} question asked may depend on the oracle's answer to question $m-1$. Formally, an n-query strategy $\mathcal{S}$ is conveniently represented as a *4-tree* of height n whose points are, when $\mathbf{A}$ is classical, occurrences of elements of $Sen(L_P)$ and, when $\mathbf{A}$ is magical, occurrences of elements of $Sen(\mathcal{L})$. A 4-tree is a tree in which each point that is not an endpoint has exactly 4 successors. In the case of a query strategy, the four successors of a point σ represent the follow-up question to σ that will be asked conditional on the answer—(0,1), (1, 0), (1, 1) or (0, 0)—received to σ. In *executing* a query strategy $\mathcal{S}$, the agent starts by asking the question corresponding to the origin of $\mathcal{S}$ and he asks further questions depending on the answers he received to previous ones. For instance, if the answer received by the agent to his first question was $(0, 1)$, he consequently asks the corresponding follow-up question. Similarly for the other answers and for questions "higher up" $\mathcal{S}$. Let us illustrate the notions just defined via an example.

Example 4.3 1-out-of-4 and a 2-query strategy of $\mathbf{A_{cla}}$
In the query structure *1-out-of-4*, an agent knows that exactly one of the propositions in $\{p_1, p_2, p_3, p_4\}$ is true and his task is to find out which one it is. Thus, 1-out-of-4 $= \langle \{\theta_1 \vee \theta_2 \vee \theta_3 \vee \theta_4\}, \{p_1, p_2, p_3, p_4\} \rangle$, where:

$$\theta_1 := p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4, \qquad \theta_2 := p_2 \wedge \neg p_1 \wedge \neg p_3 \wedge \neg p_4,$$
$$\theta_3 := p_3 \wedge \neg p_2 \wedge \neg p_1 \wedge \neg p_4, \qquad \theta_4 := p_4 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_1.$$

[6] If we identify agents with the language they speak, we have one classical agent and a lot of magical agents; one for each truth language $\mathcal{L}$.

Here is a 2-query strategy of $\mathbf{A_{cla}}$.



In executing this strategy, $\mathbf{A_{cla}}$ first asks the question $(p_1 \vee p_2)$ and then $\mathbf{A_{cla}}$ asks a follow-up question which depends on the answer that the oracle gave to $(p_1 \vee p_2)$. Depending on whether the oracle answered $(p_1 \vee p_2)$ with $(1,0)$, $(0,1)$, $(0,0)$ or $(1,1)$ the agent respectively asks p_1, p_5, $(p_7 \vee p_8)$ or $(p_1 \vee \neg p_1)$ as a follow-up question. □

Consider what happens with $\mathbf{A_{cla}}$'s knowledge if $\mathbf{A_{cla}}$ executes the query strategy $\mathcal{S}$ of Example 4.3. The oracle answers the first question $(p_1 \vee p_2)$ with $(1,0)$ just in case p_1 or p_2 is the true proposition. If $(p_1 \vee p_2)$ is answered with $(1,0)$, $\mathbf{A_{cla}}$ asks p_1 as a follow-up question. The follow-up question p_1 is answered with $(1,0)$ just in case p_1 is true. Thus, if p_1 or p_2 is the true proposition, $\mathbf{A_{cla}}$ can decide his target after an execution of $\mathcal{S}$. However, if p_3 or p_4 is the true proposition, the oracle will answer the first question $(p_1 \vee p_2)$ with $(0,1)$ and $\mathbf{A_{cla}}$ will ask p_5 as a follow up question. Clearly, the answer to p_5 reveals no information about the truth values of $\{p_1, p_2, p_3, p_4\}$. Hence, when p_3 or p_4 is the true proposition, an execution of $\mathcal{S}$ does not allow $\mathbf{A_{cla}}$ to decide his target. We say that $\mathcal{S}$ *does not solve* 1-out-of-4, because the world may be such that $\mathbf{A_{cla}}$'s knowledge update due to executing $\mathcal{S}$ does not allow $\mathbf{A_{cla}}$ to decide his target. In the next subsection we will define $\mathbf{KU}(\mathcal{S}, w) \subseteq Sen(L_P)$, which is the knowledge update of an agent due to an execution of query strategy $\mathcal{S}$ in world w. Using $\mathbf{KU}(\mathcal{S}, w)$, we can rigorously define the notion of a query strategy *solving* a query structure.

Definition 4.11 Solving a query structure
Let $\langle \mathcal{B}, \mathcal{T} \rangle$ be a query structure and let $\mathcal{S}$ be a query strategy of an arbitrary agent $\mathbf{A}$. We say that $\mathcal{S}$ *solves* $\langle \mathcal{B}, \mathcal{T} \rangle$ just in case[7] we have that, for each $w \in W$ and for each $\sigma \in \mathcal{T}$, after executing $\mathcal{S}$ in w, $\mathbf{A}$ can decide $\mathcal{T}$. Or, which is equivalent, we may say that $\mathcal{S}$ *solves* $\langle \mathcal{B}, \mathcal{T} \rangle$ just in case for every $w \in W$ we have that:

$$\forall \sigma \in \mathcal{T}\colon\ \mathcal{B} \cup \mathbf{KU}(\mathcal{S}, w) \vdash_{L_P} \sigma \text{ or } \mathcal{B} \cup \mathbf{KU}(\mathcal{S}, w) \vdash_{L_P} \neg\sigma \qquad \square$$

Clearly, by replacing, in the query strategy of Example 4.3, the follow up question p_5 with the question p_3, $\mathbf{A_{cla}}$ obtains an alternative query strategy which does solve 1-out-of-4. The *classical* and *magical query complexity* of a query structure are defined as follows.

Definition 4.12 Query complexity, classical and magical
The *classical query complexity* $\mathcal{C}_{\langle \mathcal{B}, \mathcal{T} \rangle}$ of a query structure $\langle \mathcal{B}, \mathcal{T} \rangle$ is the least n for which $\mathbf{A_{cla}}$ has a query strategy $\mathcal{S}$ which solves $\langle \mathcal{B}, \mathcal{T} \rangle$.

[7]Note that, when w is a world whose union with $\mathcal{B}$ is an inconsistent set of sentences, $\mathcal{S}$ trivially decides $\mathcal{T}$ *in that world* w.

The *magical query complexity* $\mathcal{M}_{\langle\mathcal{B},\mathcal{T}\rangle}$ of a query structure $\langle\mathcal{B},\mathcal{T}\rangle$ is the least n for which some[8] $\mathbf{A_{ma}}$ has a query strategy $\mathcal{S}$ which solves $\langle\mathcal{B},\mathcal{T}\rangle$. □

In Section 4.4.3, we will prove Theorem 4.1, which is the central result of this paper.

Theorem 4.1 The Computational Power of Self-Referential Truth Self-referential truth has computational power, as:

$$\text{There exist query structures } \langle\mathcal{B},\mathcal{T}\rangle \text{ such that: } \mathcal{M}_{\langle\mathcal{B},\mathcal{T}\rangle} < \mathcal{C}_{\langle\mathcal{B},\mathcal{T}\rangle}$$

Proof: See Section 4.4.3. □

Before we establish the *CPSRT* result we first define, in Section 4.4.2, the knowledge update $\mathbf{KU}(\mathcal{S},w)$ which an arbitrary agent $\mathbf{A}$ receives by executing $\mathcal{S}$ in w.

4.4.2 Knowledge updates from query strategies

In this section we define $\mathbf{KU}(\mathcal{S},w)$ which is the knowledge update that an arbitrary agent $\mathbf{A}$ receives by executing query strategy $\mathcal{S}$ in w. Before we go over to the actual definition of $\mathbf{KU}(\mathcal{S},w)$, we first present the *rationale* of this definition. Observe that:

1. By executing an n-query strategy $\mathcal{S}$ in w, $\mathbf{A}$ learns the answers of the oracle to n questions; say that $\mathbf{A}$ learns $\mathcal{V}^{\mathsf{as}}(\sigma_1,w),\ldots,\mathcal{V}^{\mathsf{as}}(\sigma_n,w)$.
2. Learning the assertoric value $\mathcal{V}^{\mathsf{as}}(\sigma,w)$ of σ is learning whether $O^{\mathsf{as}}_w(\mathfrak{T}^\sigma_A)$ or $C^{\mathsf{as}}_w(\mathfrak{T}^\sigma_A)$ and also learning whether $O^{\mathsf{as}}_w(\mathfrak{T}^\sigma_D)$ or $C^{\mathsf{as}}_w(\mathfrak{T}^\sigma_D)$.
3. Learning that $O^{\mathsf{as}}_w(\mathfrak{T}^\sigma_X)$ is learning that $O^{\mathsf{as}}_w(B)$ for some $B\in\mathfrak{T}^\sigma_X$.
   Learning that $C^{\mathsf{as}}_w(\mathfrak{T}^\sigma_X)$ is learning that $C^{\mathsf{as}}_w(B)$ for all $B\in\mathfrak{T}^\sigma_X$.

For some assertoric trees it is *uninformative* for $\mathbf{A}$ to be told that the tree is open or to be told that it is closed[9], the reason being that it is *a priori* that such a tree is open or closed. An assertoric tree $\mathfrak{T}^\sigma_X$ is called *a priori* just in case $\mathfrak{T}^\sigma_X$ is not *a posteriori*, where $\mathfrak{T}^\sigma_X$ is called a posteriori just in case:

$$\exists w,w'\in W : C^{\mathsf{as}}_w(\mathfrak{T}^\sigma_X) \text{ and } O^{\mathsf{as}}_{w'}(\mathfrak{T}^\sigma_X) \tag{4.10}$$

We let $O_{\mathfrak{T}^\sigma_X}\in Sen(L_P)$ and $C_{\mathfrak{T}^\sigma_X}\in Sen(L_P)$ represent the information learned by $\mathbf{A}$ when he finds out that $\mathfrak{T}^\sigma_X$ is open, respectively closed. The observation that from a priori trees nothing can be learned is modeled by letting, for any a priori tree $\mathfrak{T}^\sigma_X$, $O_{\mathfrak{T}^\sigma_X}=C_{\mathfrak{T}^\sigma_X}=\top$, where $\top := p_1\vee\neg p_1$. When an agent finds out that an a posteriori tree $\mathfrak{T}^\sigma_X$ is open or closed, he does learn something; he can extract all the information present on the *a posteriori* branches of $\mathfrak{T}^\sigma_X$, where a branch B is *a posteriori* just in case:

$$\exists w,w'\in W : C^{\mathsf{as}}_w(B) \text{ and } O^{\mathsf{as}}_{w'}(B) \tag{4.11}$$

[8] Identifying magical agents with the particular truth language they speak.

[9] For sake of brevity, we will write 'open' instead of 'open according to the assertoric closure conditions in the actual world'. Similarly for 'closed'.

Clearly, an a posteriori tree has at least one a posteriori branch.[10] With $\mathfrak{T}^{\sigma}_{X}$ an a posteriori tree, $O_{\mathfrak{T}^{\sigma}_{X}}$ and $C_{\mathfrak{T}^{\sigma}_{X}}$ can be defined in terms of $O_B, C_B \in Sen(L_P)$, where intuitively, O_B is what **A** learns when he finds out that an a posteriori branch B is open while C_B is what **A** learns when he finds out that an a posteriori branch B is closed. Once we have defined the sentences C_B and O_B we can thus complete our definition of $O_{\mathfrak{T}^{\sigma}_{X}}$ and $C_{\mathfrak{T}^{\sigma}_{X}}$ as follows[11]:

$$O_{\mathfrak{T}^{\sigma}_{X}} = \begin{cases} \bigvee\{O_B \mid B \in \mathfrak{T}^{\sigma}_{X}, B \text{ is a posteriori}\}, & \mathfrak{T}^{\sigma}_{X} \text{ is a posteriori;} \\ \top, & \mathfrak{T}^{\sigma}_{X} \text{ is a priori.} \end{cases}$$

$$C_{\mathfrak{T}^{\sigma}_{X}} = \begin{cases} \bigwedge\{C_B \mid B \in \mathfrak{T}^{\sigma}_{X}, B \text{ is a posteriori}\}, & \mathfrak{T}^{\sigma}_{X} \text{ is a posteriori;} \\ \top, & \mathfrak{T}^{\sigma}_{X} \text{ is a priori.} \end{cases}$$

If **A** finds out that an a posteriori branch B is open, he thereby finds out that all sentences of (atomic) form X_p are in agreement with the world, while if he finds out that B is closed, he thereby finds out that at least one sentence of (atomic) form X_p is in disagreement with the world. More precisely then, the sentences C_B and O_B can be defined as follows. With $p \in P$, we let $(A_p)^+ = p, (D_p)^+ = \neg p$, $(A_p)^- = \neg p$ and $(D_p)^- = p$ and we let:

$$O_B = \bigwedge\{(X_p)^+ \mid X_p \in B\}, \qquad C_B = \bigvee\{(X_p)^- \mid X_p \in B\}$$

Now that we have completed our definition of $O_{\mathfrak{T}^{\sigma}_{X}}$ and $C_{\mathfrak{T}^{\sigma}_{X}}$, we can define the information that **A** learns by learning the assertoric value of $\sigma \in Sen(\mathcal{L})$. If **A** learns that $\mathcal{V}^{\mathsf{as}}(\sigma, w) = (x, y)$, he learns that $\mathcal{U}(\sigma, (x, y)) \in Sen(L_P)$, where:

$$\mathcal{U}(\sigma,(1,0)) = O_{\mathfrak{T}^{\sigma}_{A}} \wedge C_{\mathfrak{T}^{\sigma}_{D}} \qquad \mathcal{U}(\sigma,(0,1)) = C_{\mathfrak{T}^{\sigma}_{A}} \wedge O_{\mathfrak{T}^{\sigma}_{D}}$$
$$\mathcal{U}(\sigma,(0,0)) = C_{\mathfrak{T}^{\sigma}_{A}} \wedge C_{\mathfrak{T}^{\sigma}_{D}} \qquad \mathcal{U}(\sigma,(1,1)) = O_{\mathfrak{T}^{\sigma}_{A}} \wedge O_{\mathfrak{T}^{\sigma}_{D}}$$

As by executing an n-query strategy $\mathcal{S}$ in w, **A** learns, say, $\mathcal{V}^{\mathsf{as}}(\sigma_1, w), \ldots, \mathcal{V}^{\mathsf{as}}(\sigma_n, w)$, his knowledge update $\mathbf{KU}(\mathcal{S}, w)$ due to the execution of $\mathcal{S}$ in w is defined as follows:

$$\mathbf{KU}(\mathcal{S}, w) = \{\mathcal{U}(\sigma_1, \mathcal{V}^{\mathsf{as}}(\sigma_1, w)), \ldots, \mathcal{U}(\sigma_n, \mathcal{V}^{\mathsf{as}}(\sigma_n, w))\} \tag{4.12}$$

The notions that are defined in this subsection are illustrated by means of the following example.

Example 4.4 Knowledge update (Example 1 continued)
With $\gamma := (\neg T(\lambda) \wedge p_1)$ as in Example 1, we have that:

$$\mathfrak{T}^{\gamma}_{A} = \{\{A_{(\neg T(\lambda) \wedge p_1)}, A_{\neg T(\lambda)}, A_{p_1}, D_{T(\lambda)}, D_{\neg T(\lambda)}, A_{T(\lambda)}\}\}$$
$$\mathfrak{T}^{\gamma}_{D} = \{\{D_{(\neg T(\lambda) \wedge p_1)}, D_{\neg T(\lambda)} A_{T(\lambda)}, A_{\neg T(\lambda)}, D_{T(\lambda)}\}, \{D_{(\neg T(\lambda) \wedge p_1)}, D_{p_1}\}\}$$

Naming the branches as in Example 1, we write $\mathfrak{T}^{\gamma}_{A} = \{A_1\}$ and $\mathfrak{T}^{\gamma}_{A} = \{D_1, D_2\}$. As A_1 is the only branch of $\mathfrak{T}^{\gamma}_{A}$ and as A_1 is a priori, we have that $C_{\mathfrak{T}^{\sigma}_{A}} = O_{\mathfrak{T}^{\sigma}_{A}} = \top$. $\mathfrak{T}^{\gamma}_{D}$ is an a posteriori tree which has a branch, D_1, that is a priori and a branch,

[10] Observe that the a priori tree $\mathfrak{T}^{p_1 \vee \neg p_1}_{A}$ testifies that it is not the case that all the branches of an a priori tree are a priori. However, it clearly holds that if all the branches of $\mathfrak{T}^{\sigma}_{X}$ are a priori then $\mathfrak{T}^{\sigma}_{X}$ is itself a priori.

[11] With S a finite set of sentences, $\bigvee S$ and $\bigwedge S$ denote the disjunction respectively conjunction of all the members of S.

D_2, that is a posteriori. Observe that $O_{D_2} = \neg p_1$ and that $C_{D_2} = p_1$ and so, as D_2 is the only a posteriori branch of $\mathfrak{T}_D^\gamma$, we have that $O_{\mathfrak{T}_D^\gamma} = \neg p_1$ and that $C_{\mathfrak{T}_D^\gamma} = p_1$. The function $\mathcal{U}$ is thus as follows:

$$\mathcal{U}(\gamma,(1,0)) = (\top \wedge p_1) \qquad \mathcal{U}(\gamma,(0,1)) = (\top \wedge \neg p_1)$$
$$\mathcal{U}(\gamma,(0,0)) = (\top \wedge p_1) \qquad \mathcal{U}(\gamma,(1,1)) = (\top \wedge \neg p_1)$$

We can think of γ as a 1-query strategy. With $U \subset W$ the set of all worlds u such that $p_1 \in u$ and with $V \subset W$ the set of all worlds v such that $p_1 \notin v$, we see that for all $u \in U$ and for all $v \in V$ we have that:

$$\mathbf{KU}(\gamma,u) = \{(\top \wedge p_1)\}, \qquad \mathbf{KU}(\gamma,v) = \{(\top \wedge \neg p_1)\}$$

The query strategy γ solves the query structure $\langle \varnothing, \{p_1\}\rangle$, as for all $w \in W$ we have that:

$$\varnothing \cup \mathbf{KU}(\gamma,w) \vdash_{L_P} p_1 \quad \text{or} \quad \varnothing \cup \mathbf{KU}(\gamma,w) \vdash_{L_P} \neg p_1$$

The example illustrates that, although $\mathcal{U}$ is defined for all question-answer pairs, this does not mean in general that all four answers to a question are possible answers; depending on the truth value of p_1, the oracle will either answer γ with $(0,1)$ or $(0,0)$. □

Although, as shown in Example 4.4, $\gamma := (\neg T(\lambda) \wedge p_1)$ solves the query structure $\langle \varnothing, \{p_1\}\rangle$ the solution is unnecessarily complicated, as asking p_1 also solves $\langle \varnothing, \{p_1\}\rangle$; although we solved $\langle \varnothing, \{p_1\}\rangle$ by asking a single question of $\mathcal{L}$ we could also solve the structure in a single classical query, that is by asking a single question of L_P. In the next section, we discuss a query structure whose one question solution in $\mathcal{L}$ has no one question counterpart in L_P.

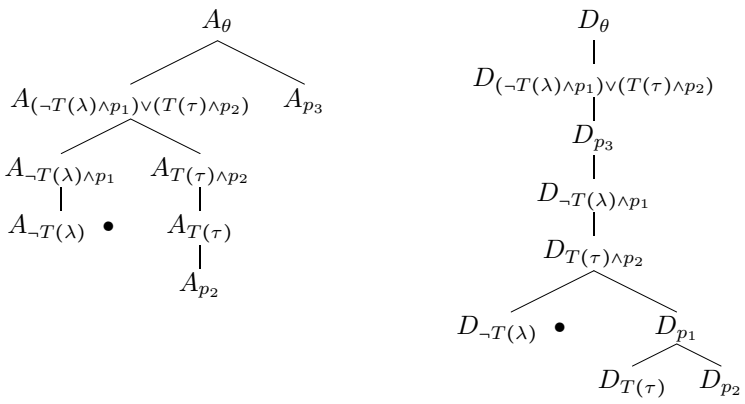
4.4.3 Magically, you can't do this classically

In this section we illustrate the computational power of self-referential truth. We do so by showing that the magical query complexity of the query structure 1-out-of-4 (see Example 4.3) is strictly less than its classical query complexity. Remember that, with θ_i as in Example 4.3, 1-out-of-4 = $\langle \{\theta_1 \vee \theta_2 \vee \theta_3 \vee \theta_4\}, \{p_1,p_2,p_3,p_4\}\rangle$, i.e. the agent's background knowledge is such that he knows that exactly 1 out of 4 given p_i is true and his target is find out which one it is. Clearly, $\mathbf{A_{cla}}$ can solve 1-out-of-4 in 2 questions and equally clearly, $\mathbf{A_{cla}}$ can not solve 1-out-of-4 in 1 question. Hence, the classical query complexity of 1-out-of-4 is 2. However, the magical query complexity of *1-out-of-4* is 1, as is illustrated by the strategy consisting of the single question θ.

$$\theta := ((\neg T(\lambda) \wedge p_1) \vee (T(\tau) \wedge p_2)) \vee p_3$$

In order to illustrate that asking θ allows $\mathbf{A_{ma}}$ to decide his target, we display $\mathfrak{T}_A^\theta$ and $\mathfrak{T}_D^\theta$.[12]

[12] To save some space, we display in fact abbreviations of those trees; a bullet (•) indicates that we do not work out the official steps after this point, as it is clear that the resulting branch(es) are a priori and closed.



Numbering the branches from left to right, we have that $\mathfrak{T}^\theta_A = \{A_1, A_2, A_3\}$ and $\mathfrak{T}^\theta_D = \{D_1, D_2, D_3\}$. Both $\mathfrak{T}^\theta_A$ and $\mathfrak{T}^\theta_D$ are a posteriori and both have a single a priori branch, respectively A_1 and D_1. With respect to the a posteriori branches, observe that:

- $O_{A_2} = p_2, \quad C_{A_2} = \neg p_2$
- $O_{A_3} = p_3, \quad C_{A_3} = \neg p_3$
- $O_{D_2} = \neg p_1 \wedge \neg p_3, \quad C_{D_2} = p_1 \vee p_3$
- $O_{D_3} = \neg p_1 \wedge \neg p_2 \wedge \neg p_3, \quad C_{D_3} = p_1 \vee p_2 \vee p_3$

And so we get that:

- $O_{\mathfrak{T}^\theta_A} = p_2 \vee p_3$
- $C_{\mathfrak{T}^\theta_A} = \neg p_2 \wedge \neg p_3$
- $O_{\mathfrak{T}^\theta_D} = (\neg p_1 \wedge \neg p_3) \vee (\neg p_1 \wedge \neg p_2 \wedge \neg p_3)$
- $C_{\mathfrak{T}^\theta_A} = (p_1 \vee p_3) \wedge (p_1 \vee p_2 \vee p_3)$

And so the function $\mathcal{U}$ has the following properties:

(x, y)	$\mathcal{U}(\theta, (x, y))$
$(0,1)$	$(\neg p_2 \wedge \neg p_3) \wedge ((\neg p_1 \wedge \neg p_3) \vee (\neg p_1 \wedge \neg p_2 \wedge \neg p_3))$
$(0,0)$	$(\neg p_2 \wedge \neg p_3) \wedge ((p_1 \vee p_3) \wedge (p_1 \vee p_2 \vee p_3))$
$(1,0)$	$(p_2 \vee p_3) \wedge ((p_1 \vee p_3) \wedge (p_1 \vee p_2 \vee p_3))$
$(1,1)$	$(p_2 \vee p_3) \wedge ((\neg p_1 \wedge \neg p_3) \vee (\neg p_1 \wedge \neg p_2 \wedge \neg p_3))$

With $i \in \{1,2,3,4\}$ we let $W_i \subset W$ be the set of those worlds w_i for which $p_i \in w_i$ and such that for all $j \in \{1,2,3,4\}$, $j \neq i$ we have that $p_j \notin w_i$. With w_i an arbitrary world in W_i, we have that:

w_i	$\mathbf{KU}(\theta, w_i)$
w_1	$\{\mathcal{U}(\theta, (0,0))\}$
w_2	$\{\mathcal{U}(\theta, (1,1))\}$
w_3	$\{\mathcal{U}(\theta, (1,0))\}$
w_4	$\{\mathcal{U}(\theta, (0,1))\}$

It is left to the reader to check that, with $\mathcal{B}$ the background knowledge of $\mathbf{A_{ma}}$ in 1-out-of-4, with $i, j \in \{1,2,3,4\}, j \neq i$ and with $W^* = W - (W_1 \cup W_2 \cup W_3 \cup W_4)$ we have that:

$$w \in W_i \Rightarrow \mathcal{B} \cup \mathbf{KU}(\theta, w) \vdash_{L_P} p_i \text{ and } \mathcal{B} \cup \mathbf{KU}(\theta, w) \vdash_{L_P} \neg p_j$$
$$w \in W^* \Rightarrow \mathcal{B} \cup \mathbf{KU}(\theta, w) \vdash_{L_P} \sigma \text{ for all } \sigma \in Sen(L_P)$$

And so, the single query strategy θ solves 1-out-of-4, establishing that the magical query complexity is strictly less than its classical query complexity, i.e. establishing the Computational Power of Self-Referential Truth. In the next section, the significance of the *CPSRT* result will be discussed. We conclude this section with the following observations.

1. 1-out-of-4* $= \langle \varnothing, \{(p_1 \wedge \neg p_2), (\neg p_1 \wedge p_2), (\neg p_1 \wedge \neg p_2), (p_1 \wedge p_2)\}\rangle$ can be solved by asking the single question $\theta^* := (\neg T(\lambda) \wedge p_1) \vee (T(\tau) \wedge p_2)$.

2. Analogous to 1-out-of-4, one can define the query structure 1-out-of-n. With $n \geq 1$, the magical query complexity of 1-out-of-4^n is n, whereas its classical query complexity is $2n$.

3. The question θ which we used to solve 1-out-of-4 exploits two non-quotational constants: λ and τ. It is also possible to solve 1-out-of-4 in a single question which uses only one non-quotational constant, as is testified by letting:

   $$\pi(c) = ((\neg T(c) \wedge p_1) \vee (T(c) \wedge p_2)) \vee p_3$$

   The reader may verify that the query strategy consisting of the single question $((\neg T(c) \wedge p_1) \vee (T(c) \wedge p_2)) \vee p_3$ solves 1-out-of-4.

4.5 Remarks on the significance of *CPSRT*

4.5.1 What does *CPSRT* have to do with computation?

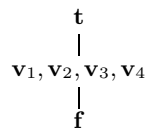
In this section I will argue against a possible objection to my presentation of Theorem 4.1 as a result about computation. The objection is, roughly, as follows. Granted, there clearly are query structures with a lower magical than classical query complexity, as indicated in Theorem 4.1. But it is misleading to speak of this result in terms of *computational* power. For the notion of magical query complexity, as defined in Definition 4.12, is not a natural notion of query (and hence computational) complexity.

In fact, I will discuss and respond to two more concrete objections, the spirit of both of which is that the notion of magical query complexity is not a natural notion of computational complexity. The two objections, the arguments for which are given below, are as follows:

1. It can be shown that, according to the *rationale* of the notion of magical query complexity, it is possible to solve, for any $n \in \mathbb{N}$, 1-out-of-n in a single question. Clearly, a notion of query complexity with such a *rationale* is not a natural notion.

2. To answer the question 'what is a natural notion of computational (query) complexity?' we should consult computer science. The query complexity of 1-out-of-4 is not equal to 1 according to the "computer scientific notion of query complexity", showing that the notion of magical query complexity is not a natural notion.

Argument for objection 1. The magical query complexity of a query structure is the least number of sentences (questions) that you have to ask an oracle in order to solve the structure, whereas the answer given by the oracle to a sentence σ is the semantic (assertoric) value of σ. Following this line of thought, it is easy to solve, for each $n \in \mathbb{N}$, 1-out-of-n in a single question. We will illustrate how this can be done for $n = 6$. Consider the structure $SIX_{\sqsubseteq} = \langle\{\mathbf{t}, \mathbf{f}, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}, \leq\rangle$ in which the order $\sqsubseteq$ partially orders the carrier SIX according to the following Hasse diagram.



Hence $\mathbf{t}$ is the largest$_{\sqsubseteq}$ element of SIX, $\mathbf{f}$ is its smallest$_{\sqsubseteq}$ element and the $\mathbf{v}_i$'s are incomparable with one another. Let L be a language which has $\wedge$ and $\vee$ as logical symbols for conjunction and disjunction and whose non-logical symbolism consists of the set of propositional atoms P and of the set $\Sigma = \{\sigma_{\mathbf{v}_1}, \sigma_{\mathbf{v}_2}, \sigma_{\mathbf{v}_3}, \sigma_{\mathbf{v}_4}\}$ of non-propositional atoms. $Sen(L)$ is the smallest set containing P and Σ and such that if $\alpha, \beta \in Sen(L)$ then $(\alpha \wedge \beta), (\alpha \vee \beta) \in Sen(L)$.

An *atomic L valuation V_A for L* is a function $V_A : P \cup \Sigma \rightarrow SIX$ which satisfies $V_A(\sigma_{\mathbf{v}_i}) = \mathbf{v}_i$ for all $\sigma_{\mathbf{v}_i} \in \Sigma$ and $V_A(p) \in \{\mathbf{t}, \mathbf{f}\}$ for all $p \in P$. A valuation function V for L is a function $V : Sen(L) \rightarrow SIX$ extending an atomic L valuation function V_A in a $SIX_{\sqsubseteq}$ *compositional* way, i.e V is such that:

- $V(\alpha) = V_A(\alpha)$ for all $\alpha \in P \cup \Sigma$ and some V_A.
- $V(\alpha \vee \beta) = sup_{\sqsubseteq}(\{V(\alpha), V(\beta)\})$
- $V(\alpha \wedge \beta) = inf_{\sqsubseteq}(\{V(\alpha), V(\beta)\})$

Assume that we have available an oracle whose answer to $\sigma \in Sen(L)$ is equal to $V(\sigma)$, for some L valuation function V, and consider the query structure 1-out-of-6 $= \langle\mathcal{B}, \{p_1, p_2, p_3, p_4, p_5, p_6\}\rangle$, where $\mathcal{B}$ formalizes that the agent knows that exactly one of the six p_i is true ($\mathbf{t}$) and that all other p_i are false ($\mathbf{f}$). Now ask the following question to the oracle.

$$\vartheta := (\sigma_{\mathbf{v}_1} \wedge p_1) \vee (\sigma_{\mathbf{v}_2} \wedge p_2) \vee (\sigma_{\mathbf{v}_3} \wedge p_3) \vee (\sigma_{\mathbf{v}_4} \wedge p_4) \vee p_5$$

As is easily observed by an inspection of the Hasse diagram, taking the background knowledge $\mathcal{B}$ into consideration, we have that:

- $V(p_i) = \mathbf{t} \Leftrightarrow V(\vartheta) = \mathbf{v}_i$ for $i \in \{1, 2, 3, 4\}$
- $V(p_5) = \mathbf{t} \Leftrightarrow V(\vartheta) = \mathbf{t}$
- $V(p_6) = \mathbf{t} \Leftrightarrow V(\vartheta) = \mathbf{f}$

Hence, using the same assumptions as those underlying the notion of magical query complexity, we can solve 1-out-of-6 in 1 question. In fact, by an obvious extension of the argument just given we can solve, for arbitrary n, the query structure 1-out-of-n in a single question. This shows that the (rationale of) the notion of magical complexity is not natural.

Response to objection 1. The algebraic argument given above is certainly correct. However, the mathematical possibility of coming up with such an L and V is no argument against the notion of magical query complexity. Assertoric semantics, used to define $\mathcal{V}^{\mathsf{as}}$ which solves 1-out-of-4 in 1 question is, so I claim, a *natural* semantics, arising out of the rules that govern *our* practices of asserting and denying sentences. The fact that one can generalize the algebraic structure of this semantics and come up with a notion of query complexity according to which 1-out-of-6 can be solved in a single question does not show that *we* can solve 1-out-of-6 in a single question. I realize that this defense against objection 1 commits me to the assertion that *we can* solve 1-out-of-4 in 1 question, which confronts me directly with objection 2.

Argument for objection 2. The unit of computation is the *bit* and accordingly, the smallest question you can ask to an oracle, thought of as a computational entity, is a 1 bit question on which the oracle answers with either 1 or 0. A computational model of the query structure 1-out-of-4 roughly has the following form. Coding the background knowledge of the agent **A** as states of bits we get that **A** knows, of an unknown bits state (x, y), that it is either $(0, 0), (0, 1), (1, 0)$ or $(1, 1)$. The oracle has a register containing the actual value of (x, y). A query to the oracle is a 1 bit question about this register. For instance, **A** can ask $x = 1 \wedge \neg y = 0$ on which the oracle answers with 1 if and only if $(x, y) = (1, 1)$ and with 0 otherwise. Clearly, on the (classical) notion of query complexity associated with this computational model it is not possible to solve 1-out-of-4 in a single question. Hence, the notion of magical query complexity, according to which it is possible, is not a natural notion.

Response to objection 2. The thought that the unit of computation is the bit is outdated. In *quantum computation*, the unit of computation is not the bit but rather the *qubit*. Researchers in the area of quantum computation have defined the notion of *quantum query complexity*[13] which, according to them, is the most natural notion of query complexity in the quantum computational framework. The query structure 1-out-of-4 can also be modeled in the quantum computational framework. Interestingly, *1-out-of-4 can be solved in a single quantum query*. The algorithm which realizes this speed up over classical computation is known as *Grover's search algorithm* ([22]). An accessible introduction to quantum computation, containing a nice description of Grover's search algorithm, is the textbook ([39]). And so, as there is a natural notion of query complexity according to which 1-out-of-4 can be solved in a single question, objection 2 is not a legitimate objection. Now the question arises how and in what sense assertoric semantics is related to quantum computation. Currently, the author is exploring this interesting question. However, addressing this question in any detail is far beyond the scope of this paper.

We addressed two possible objections against our presentation of the notion of magical query complexity as a certain measure of computational complexity.

[13] For an overview of various notions of quantum computational complexity, amongst which quantum query complexity, see ([12]).

An interesting question is how the notion of magical query complexity is related to other measures of computational complexity such as *time* or (memory) *space*. We showed that the availability of self-referential resources decreased the query complexity of certain computational problems. However, the decrease of the complexity on the query complexity scale may be counterbalanced by an increase of the complexity according to other complexity measures[14]. For instance, although the availability of self-referential resources ensures that we need less queries to solve certain query problems, it may very well be that the oracle (computer) needs more time and / or memory space to answer our (self-referential) queries. A related point is that the agent needs to translate the answers of the oracle to his queries into information about the query problem under consideration: it may very well be that translating the answers to self-referenial queries consumes more time and / or memory space than translating the answers to classical queries. I hope to explore such issues, concerning the relation between self-referential resources and distinct measures of computational complexity, in future work.

4.5.2 *CPSRT* and deflationism: friends or foes?

Claiming that (self-referential) truth has computational power seems diametrically opposed to the nowadays predominant *deflationary* accounts of truth, according to which truth is an insubstantial and light notion. I do think that this paper's results can be conjoined with philosophical argumentation to point out that an important deflationary account of truth, the *minimalist conception* of truth as defended by Horwich, is wrong. The central claim of Horwich's position is illustrated by the following quote.

> The entire conceptual and theoretical role of truth may be explained on the basis of all uncontroversial instances of the equivalence schema: (E) It is true that p if and only if p.
>
> (Horwich, [28, p5])

By uncontroversial instances of the equivalence schema Horwich explicitly excludes the leading actors of this paper: Liars and Truthtellers. If one is willing to admit that truth has computational power, it seems that truth plays a role which cannot be accounted for on the basis of all uncontroversial instances of the equivalence schema (E), but rather, that we have to appeal to the *inference rules* of our truth predicate. The *CPSRT* thus seems incompatible with a prominent deflationary account of truth: Horwich's *minimalism*.

However, this is not to say that the *CPSRT* is incompatible with all deflationary accounts of truth. The results of this paper derive from the *inferential properties* of the notion of truth—cast in the assertoric rules for the truth predicate— and an interesting question is how this paper's results are to be interpreted in light of Horsten's *inferential* deflationism ([27]), a position that thinks of truth as *essentially* an inferential notion. Inferential deflationsim acknowledges that the notion of truth is non-conservative over mathematics and also that truth 'plays a more substantial role in certain philosophical debates than at first sight might be expected'[15]. Thus, this position acknowledges the

[14] I owe this point to an anonymous referee.

[15] As is illustrated in ([27]) via a formulation of Fitch's argument in first order (modal) logic.

mathematical and *philosophical power* of truth. Still the position is qualified as a *deflationary* account of truth.

> Thus there is a deep and important sense in which truth is a light notion. This sense is captured by the thesis that truth is a property that cannot be described in terms of unrestricted general laws; there are only restricted laws of truth.
>
> (Horsten, [27, p578])

If inferential deflationism is correct, we have a deflationary account of truth which acknowledges that truth has mathematical and philosophical power. As this paper pointed out, truth has also *computational power*, a phenomenon that, in the author's opinion, has to be accounted for by any account of truth. Inferential deflationism seems an apt way to do so.

4.5.3 The Useless Liar Conviction is false

I hope that the sketch of the framework of assertoric semantics in Section 2 suffices to convince the reader that assertoric semantics is an interesting and promising approach to (self-referential) truth. Assertoric semantics, with its insistence on inference, differs both conceptually and technically from the Kripkean fixed point semantics which is the predominant way of giving a semantics for a language of self-referential truth in the literature. In particular, I consider Assertoric semantics to be an interesting alternative to Kripke's fixed point semantics.

However, independently of the status of assertoric semantics as an account of truth, I take it that the derivation of the *CPSRT* result in section 4.4, and the discussion of this result in subsections 4.5.1 and 4.5.2, suffices to show that the *CPSRT* is something which has to be accounted for by any theory of truth. In particular, I take it that the *CPSRT* rules out certain deflationary accounts of truth. To fit in with the introduction of this paper, I think that the results obtained in this paper discredit the widely shared Useless Liar Conviction. Here is a quote from Grover—the philosopher, not the discoverer of the quantum search algorithm—illustrating a typical *ULC* attitude:

> [...] I give my reasons for holding the liar *is not* a sentence that is used in a *communicatively significant* way. This means that though the liar is syntactically well-formed, and its individual words have dictionary meaning, the liar *does not have "operative meaning"*. It is a sentence with *limited philosophical interest*. I recommend that our reaction to the liar should be *similar to our reaction to* 6/0. (Grover, [21, p178], my italics)

I hope to have convinced you that the Liar *is* a sentence that can be used in a *communicatively significant way* and that therefore the Liar *has "operative meaning"* and is of *great philosophical interest*. I recommend that our reaction to the Liar should be *similar to our reaction to* $\sqrt{-1}$: let us try to come up with a non-classical space in which we can make sense of the cognitive operations that are involved in our reflections on the Liar. If we succeed, it may be possible to assert that Grover is wrong because Grover is right.

Appendix

In order to prove Proposition 4.1 we first define the notion of the *degree* of an assertoric sentence.

Definition 4.13 The degree of X_σ.
The degree $d(X_\sigma)$ of an assertoric sentence X_σ is equal to $d(\sigma)$, where $d(\sigma)$ is defined as follows.

1. $\sigma \in P$ or σ has form $T(c)$ with $c \in C \Rightarrow d(\sigma) = 0$
2. σ has form $(\alpha \vee \beta)$ or $(\alpha \wedge \beta) \Rightarrow d(\sigma) = d(\alpha) + d(\beta) + 1$
3. σ has form $\neg\alpha \Rightarrow d(\sigma) = d(\alpha) + 1$
4. σ has form $T([\alpha]) \Rightarrow d(\sigma) = d(\alpha) + 1$ □

Proposition 1 $\mathfrak{T}_X^\sigma$ is a finite set whose elements are finite sets.
Proof: The claim follows from the observation that for all assertoric rules other than those for the truth predicate , an application of a rule only gives rise to sentences of lower degree. The application of the assertoric rule for the truth predicate to an assertoric sentence $X_{T(c)}$, with $c \in C$, may lead to a sentence of greater degree or to a sentence of the same degree. However, due to the finiteness of C an increase in degree can only occur a finite number of times and hence, each assertoric tree is finite. □

Chapter 5

From Closure Games to Generalized Strong Kleene Theories of Truth

5.1 Abstract

In this paper, we study *the method of closure games*, which is a game theoretic valuation method for languages of self-referential truth, developed by the author. We prove two theorems which jointly establish that the method of closure games characterizes all 3- and 4-valued Strong Kleene theories of truth (SK theories) in a uniform manner. Another theorem states conditions under which SK theories can be combined into *Generalized Strong Kleene theories of truth* (GSK theories). In contrast to a SK theory, a GSK theory recognizes more than one sense of strong assertibility—where a sentence is strongly assertible just in case it is assertible and its negation is not. Exploiting the relations between SK theories laid bare by the method of closure games, we then show how to define 5-, 6-, 7-, 8- and 10-valued GSK theories.

5.2 Introduction

5.2.1 The method of closure games

By a *theory of truth* **T**, we mean...

> ...a theory that purports to explain for a first-order language L_T what sentences are assertible in a [ground] model M.
>
> (Gupta [23, p19])

The *method of fixed point constructions*, as developed by Kripke [33] and *the method of revision sequences* as develped by Gupta & Belnap [24], we call *frameworks* for truth. Relative to the specification of certain (framework dependent) conditions, a framework for truth defines a theory of truth. The method of fixed point constructions defines a theory of truth upon the specification of a *monotonic valuation schema* and a set of sentences which is *sound* with respect

to that schema, whereas the method of revision sequences defines a theory of truth upon the specification of a *rule of revision*. The *Method of Closure Games* (MCG) is another framework for truth, which defines theories of truth by playing (closure) games. The framework dependent conditions of MCG are so called *closure conditions*, which intuitively can be thought of as representing assertoric norms. Below, this central notion is explained in more detail.

Closure games are governed by the *assertoric rules* of L_T, examples of which are given below.

T	$\frac{A_{T(\overline{\sigma})}}{A_\sigma}$	$\frac{D_{T(\overline{\sigma})}}{D_\sigma}$

$\wedge$	$\frac{A_{(\alpha\wedge\beta)}}{A_\alpha, A_\beta}$	$\frac{D_{(\alpha\wedge\beta)}}{D_\alpha \mid D_\beta}$

$\neg$	$\frac{A_{\neg\sigma}}{D_\sigma}$	$\frac{D_{\neg\sigma}}{A_\sigma}$

Here, A_σ and D_σ stand for (a commitment to) an Assertion and Denial of σ. *As a framework* for theories of truth, MCG is not committed to a particular interpretation of the assertoric rules. However, in this paper we will only be concerned with MCG's definition of various (three and four valued) Strong Kleene theories of truth (SK theories). Under a Strong Kleene interpretation, the rules for assertion and the rules for denial both receive an "iff reading". For instance[1]:

$$\alpha \wedge \beta \text{ is assertible iff } \alpha \text{ is assertible and } \beta \text{ is assertible.} \tag{5.1}$$

$$\alpha \wedge \beta \text{ is deniable iff } \alpha \text{ is deniable or } \beta \text{ is deniable.} \tag{5.2}$$

The assertoric rules resemble the rules of a signed tableau calculus for first order logic, as studied in Smullyan [50], with rules for the truth predicate added to it. Importantly however, the assertoric rules are not used as a proof system, but rather as a semantic valuation method.

In a *closure game* for A_σ (D_σ), player $\sqcup$, who controls all signed sentences of disjunctive type (e.g., $D_{\alpha\wedge\beta}$) tries to argue that σ is assertible (deniable), while player $\sqcap$, who controls all sentences of conjunctive type (e.g. $A_{\alpha\wedge\beta}$) tries to prove player $\sqcup$ to be wrong. Whether or not $\sqcup$ is successful depends, amongst others, on the *assertoric norm* under consideration. An assertoric norm is formally represented as a *closure condition*, which is a bipartition of the set of possible *expansions* that the players may induce by picking their strategies, as explained below.

A *strategy* of a player is a mapping of each AD sentence X_σ—where $X \in \{A, D\}$—that is in his control to exactly one of the *immediate successors* of X_σ, as specified by the assertoric rule applicable to X_σ. A few examples suffice to illustrate the notion of a strategy. The immediate successors of $A_{\alpha\wedge\beta}$ are A_α and A_β and, as $A_{\alpha\wedge\beta}$ is of conjunctive type, a strategy of player $\sqcap$ maps $A_{\alpha\wedge\beta}$ to either A_α or A_β. As $A_{T(\overline{\sigma})}$ has only one immediate successor, A_σ, every strategy of player $\sqcap$ must map $A_{T(\overline{\sigma})}$ to A_σ. A strategy for player $\sqcup$, who controls $D_{\alpha\wedge\beta}$, maps $D_{\alpha\wedge\beta}$ to either D_α or D_β.

With f a strategy for player $\sqcup$, g a strategy for player $\sqcap$ and with X_σ an arbitrary AD sentence, the tuple (X_σ, f, g) defines an *expansion of* X_σ, denoted

[1]Where 'σ is assertible' is shorthand for 'it is possible to live up to the commitments involved with an assertion of σ'. Whether or not it is possible to do so depends on (the logical form of σ and on) the assertoric norms under consideration, which are formally modeled as closure conditions, as explained below. Similarly for the phrase 'σ is deniable'.

$\mathsf{exp}(X_\sigma, f, g)$. In general, an expansion of X_σ is an infinite[2] sequence of *AD* sentences whose first element is X_σ and whose successor relation respects the assertoric rules. As an example, here is the expansion of $A_{\neg T(\lambda)}$, i.e., of an assertion of the Liar[3].

$$A_{\neg T(\lambda)}, D_{T(\lambda)}, D_{\neg T(\lambda)}, A_{T(\lambda)}, A_{\neg T(\lambda)} \dots \tag{5.3}$$

Indeed, $A_{\neg T(\lambda)}$ has only one expansion and so, in *the closure game for* $A_{\neg T(\lambda)}$, none of the players can influence the expansion of $A_{\neg T(\lambda)}$ that is realized. In general, an *AD* sentence X_σ may have (infinitely) many expansions, each of which is realized by some strategy pair (f, g) of our players. For instance, $A_{P(c_1)\wedge P(c_2)}$, where $P(c_1)$ and $P(c_2)$ are atomic sentences of L, has two expansions and, in *the closure game for* $A_{P(c_1)\wedge P(c_2)}$, player $\sqcap$ can determine which one is realized. By setting $g(A_{P(c_1)\wedge P(c_2)}) = A_{P(c_1)}$, player $\sqcap$ ensures that expansion (5.4) is realized, while $g(A_{P(c_1)\wedge P(c_2)}) = A_{P(c_2)}$ realizes expansion (5.5).

$$A_{P(c_1)\wedge P(c_2)}, A_{P(c_1)}, A_{P(c_1)}, A_{P(c_1)}, \dots \tag{5.4}$$

$$A_{P(c_1)\wedge P(c_2)}, A_{P(c_2)}, A_{P(c_2)}, A_{P(c_2)}, \dots \tag{5.5}$$

Indeed, to get a uniform definition of the notion of an expansion, we assume that whenever we "hit" a signed atomic sentence of L, the expansion continues by repeating that *AD* sentence indefinitely.

A *closure condition* $\dagger(M) = \{O^\dagger_M, C^\dagger_M\}$ is a biparition of the set of all expansions into the sets $O^\dagger_M$ and $C^\dagger_M$, consisting of all open and all closed expansions in M respectively. In a closure game for X_σ played relative to $\dagger(M)$, player $\sqcup$ tries to pick his strategy f in such a way that the expansion of X_σ that is realized will be contained in $O^\dagger_M$. We will write $O^\dagger_M(X_\sigma)$, and say that X_σ *is open relative to* $\dagger(M)$, to indicate that player $\sqcup$ has a strategy which *ensures* that the expansion of X_σ ends up in $O^\dagger_M$. That is:

$$O^\dagger_M(X_\sigma) \Leftrightarrow \exists f \forall g : \mathsf{exp}(X_\sigma, f, g) \in O^\dagger_M \tag{5.6}$$

X_σ is *closed relative to* $\dagger(M)$, denoted $C^\dagger_M(X_\sigma)$, just in case not $O^\dagger_M(X_\sigma)$. As specified by (5.6), a closure condition for expansions induces a closure condition for *AD* sentences. The closure condition for *AD* sentences is used to induce a valuation for L_T, denoted $\mathcal{V}^\dagger_M$:

$$\mathcal{V}^\dagger_M(\sigma) = \begin{cases} \mathbf{a} := (1,0), & O^\dagger_M(A_\sigma) \text{ and } C^\dagger_M(D_\sigma); \\ \mathbf{b} := (1,1), & O^\dagger_M(A_\sigma) \text{ and } O^\dagger_M(D_\sigma); \\ \mathbf{n} := (0,0), & C^\dagger_M(A_\sigma) \text{ and } C^\dagger_M(D_\sigma); \\ \mathbf{d} := (0,1), & C^\dagger_M(A_\sigma) \text{ and } O^\dagger_M(D_\sigma). \end{cases} \tag{5.7}$$

In general $\mathcal{V}^\dagger_M$ may, but need not, have a range of four values. The values $\mathbf{a}$, $\mathbf{b}$, $\mathbf{n}$ and $\mathbf{d}$ are, intuitively, interpreted as "only assertible", "both assertible and deniable", "neither assertible nor deniable" and "only deniable" respectively. In a little more detail, $\mathcal{V}^\dagger_M(\sigma) = \mathbf{a}$ indicates that it is allowed to assert, but not to deny, sentence σ in ground model M *according to the norms for assertion and*

[2]Below, we explain how an expansion continues when it "hits" a signed atomic sentence of L.

[3]As before, λ denotes $\neg T(\lambda)$.

denial that are specified by †.

So, besides our semantic— in contrast to a proof theoretic—use of the assertoric rules, another distinguishing feature with respect to the typical use of "signed tableau" rules is that the notion of an expansion, and not that of a *branch*, is at the heart of the MCG. As we will see in Section 5.4, where we state our two *stable judgement theorems*, it is the notion of an expansion which gives the MCG the means to characterize all 3- and 4-valued SK theories in a uniform manner. In Section 5.5, we investigate a variant of the MCG which is formulated in terms of branches. As we will see, this variant allows us to capture Kripke's 4-valued "modal theory of truth" $\mathcal{K}^{\mathbf{4}}$, which he defined in [33] by quantifying over all 3-valued SK theories[4].

5.2.2 (Generalized) Strong Kleene theories of truth

In this paper, we introduce the notion of a *Generalized Strong Kleene theory of truth*, or *GSK theory*, which generalizes the algebraic characterization of a SK theory in such a way that it becomes applicable to theories of truth which recognize more than four semantic values. In contrast to a SK theory, a GSK theory may recognize more than one sense of strong assertibility—where a sentence is strongly assertible just in case it is assertible and its negation is not. As an example of a GSK theory, let us consider the theory $\mathbb{V}^{\mathbf{8}+}$, an eight valued GSK theory that will be defined later on in this paper. The eight semantic values of $\mathbb{V}^{\mathbf{8}+}$ are given by the set $\{\mathbf{a}_g, \mathbf{a}_i, \mathbf{a}_e, \mathbf{b}_e, \mathbf{n}_e, \mathbf{d}_e, \mathbf{d}_i, \mathbf{d}_g\}$. Six of the semantic values come in pairs: $(\mathbf{a}_g, \mathbf{d}_g)$, $(\mathbf{a}_i, \mathbf{d}_i)$ and $(\mathbf{a}_e, \mathbf{d}_e)$. We call these three pairs *strong assertoric pairs*, the reason being that according to $\mathbb{V}^{\mathbf{8}+}$, negation interchanges the elements within such a pair. For instance: $\mathbb{V}^{\mathbf{8}+}_M(\neg\sigma) = \mathbf{a}_e \Leftrightarrow \mathbb{V}^{\mathbf{8}+}_M(\sigma) = \mathbf{d}_e$. The crucial distinction between a SK theory and a GSK theory, is as follows: *according to a GSK theory, there may be more than one strong assertoric pair*. In fact, this is the only distinction between SK and GSK theories, as we may illustrate via the lattice of $\mathbb{V}^{\mathbf{8}+}$:

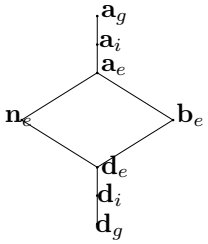


Figure 5.1: Hasse diagram of $\mathbf{8}^+_{\leq}$, the lattice of $\mathbb{V}^{\mathbf{8}+}$.

Besides interchanging the members of the three strong assertoric pairs, negation acts as the identity on $\mathbf{b}_e$ and $\mathbf{n}_e$. Further, conjunction and disjunction act as meet and join in the lattice $\mathbf{8}^+_{\leq}$, and universal and existential quantification act as generalized conjunction and disjunction. The interpretation of the eight semantic values of $\mathbb{V}^{\mathbf{8}+}$ is as follows:

[4]For instance, $\mathcal{K}^{\mathbf{4}}$ valuates the Liar as $\mathbf{n}$, as there is no 3-valued SK theory in which it is valuated as $\mathbf{a}$ and also, there is no 3-valued SK theory in which it is valuated as $\mathbf{d}$.

$\mathbf{a}_g$: grounded, only assertible.
$\mathbf{a}_i$: ungrounded, intrinsic, only assertible.
$\mathbf{a}_e$: ungrounded, extrinsic, only assertible.
$\mathbf{b}_e$: ungrounded, extrinsic, both assertible and deniable.
$\mathbf{n}_e$: ungrounded, extrinsic, neither assertible nor deniable.
$\mathbf{d}_e$: ungrounded, extrinsic, only deniable.
$\mathbf{d}_i$: ungrounded, intrinsic, only deniable.
$\mathbf{d}_g$: grounded, only deniable.

The grounded/ungrounded distinction derives from Kripke's *Strong Kleene minimal fixed point theory*, which I will denote by $\mathcal{K}$, while the intrinsic/extrinsic distinction derives from Kripke's *Strong Kleene maximal intrinsic fixed point theory*, denoted by $\mathcal{K}^+$. Although $\mathcal{K}$ and $\mathcal{K}^+$ are familiar theories, we feel that their definition via the MCG (given in Section 3 and 4 respectively) sheds some illuminating new light on these theories. The assertoric distinctions $(\mathbf{a}, \mathbf{b}, \mathbf{n}, \mathbf{d})$ derive from a novel 4-valued SK theory that will be defined, in Section 5.5, using the MCG. $\mathbb{V}^{\mathbf{8}+}$, and all other GSK theories that will be defined in this paper, are obtained as a combination of certain SK theories. The conditions under which a combination of 3- and 4-valued SK theories can be turned into a GSK theory are formulated by the *assertoric transfer theorem* (Section 5.3, Theorem 5.1). A more detailed account of the interpretation of $\mathbb{V}^{\mathbf{8}+}$ is postponed to the last section.

5.2.3 Structure of the paper

The structure of this paper is as follows. Section 5.3 gives some general preliminaries, defines the notion of a GSK theory precisely and contains the assertoric transfer theorem (Theorem 5.1). Section 5.4 presents the MCG in more detail and there we prove the first and second stable judgement theorem (Theorem 5.2 and 5.3 respectively). In Section 5.5 we show how to define Kripke's "modal theory of truth" $\mathcal{K}^{\mathbf{4}}$ via a variant of the MCG that trades in the notion of an expansion for that of a branch. Further, we use our representation of $\mathcal{K}^{\mathbf{4}}$ to define closure conditions which induce Kripke's maximal intrinsic fixed point, $\mathcal{K}^+$, via the MCG. Section 5.6 is devoted to the definition of various 5-, 6-, 7-, 8- and 10-valued GSK theories. Section 5.7 concludes. The paper contains two appendices. Appendix I contains a proof of a proposition appearing in Section 5.4, and Appendix II discusses *Yablo's paradox* (cf. Yablo [69]) in terms of the MCG.

5.3 Theories of truth and ground models

L_T will denote a first order language without function symbols, with *identity* ($\approx$), a *truth predicate* (T) and with a *quotational name* ($[\sigma]$) for each sentence σ of L_T. L will denote the language that is exactly like L_T, except for the fact that it does not contain the truth predicate T. A *ground model* $M = (D, I)$ is an interpretation of L such that $Sen(L_T) \subseteq D$ and such that $I([\sigma]) = \sigma$ for all $\sigma \in Sen(L_T)$. A sentence may be denoted in various ways; $\overline{\sigma}$ will be used to denote any closed term, quotational name or not, which denotes σ in M. We will make the simplifying assumption that, given a ground model

$M = (D, I)$, there is, for each of the members of its domain, a constant symbol in the language which refers to that element. This assumption has the advantage that quantification can be treated substitutionally, so that we do not need to be bothered with variable assignments. With respect to $Sen(L_T) \subseteq D$ this assumption is unnecessary, as every sentence contains, by definition, at least one name: its quotational name. As L_T is assumed not to contain function symbols, all the closed terms of L_T are given by its set of constant symbols, which will be denoted by $Con(L_T)$. Observe that $[\forall x T(x)] \approx [\forall x T(x)]$ is guaranteed to be a sentence of L_T. Given a ground model M, $\mathcal{C}_M : Sen(L) \to \{\mathbf{a}, \mathbf{d}\}$ denotes the *classical valuation* of L based on M and is defined as usual[5]. Note that $\mathcal{C}_M([\forall x T(x)] \approx [\forall x T(x)]) = \mathbf{a}$ and $\mathcal{C}_M([\forall x T(x)] \approx [\exists x T(x)]) = \mathbf{d}$ for any ground model M. A *theory of truth* $\mathbf{T}$ takes a ground model M as input and outputs a semantic valuation $\mathbf{T}_M$ of the sentences of L_T. That is, $\mathbf{T}$ outputs a function $\mathbf{T}_M : Sen(L_T) \to \mathbf{V}$, where $\mathbf{V}$ contains *the semantic values recognized*[6] *by* $\mathbf{T}$. With $\mathbf{T}$ a theory of truth, $\top_{\mathbf{T}} = \mathbf{T}_M([\forall x T(x)] \approx [\forall x T(x)])$ and $\bot_{\mathbf{T}} = \mathbf{T}_M([\forall x T(x)] \approx [\exists x T(x)])$ are called the *classical top value* and *classical bottom value* of $\mathbf{T}$ respectively. Not any semantic valuation of the sentences of L_T qualifies as the valuation of a theory of truth. In this paper, we assume that in order for $\mathbf{T}$ to qualify as a theory of truth, $\mathbf{T}_M$ should *respect the world* and the *identity of truth*, as defined below. Besides these two familiar conditions we impose one further, arbitrary but technically convenient, condition: every truth ascription to an object which is not a sentence is to be valuated as $\bot_{\mathbf{T}}$ by a theory of truth $\mathbf{T}$.

Definition 5.1 Theory of truth
Let $\mathbf{T}$ be a valuation method which, given a ground model $M = (D, I)$, outputs a valuation function $\mathbf{T}_M : Sen(L_T) \to \mathbf{V}$. We say that $\mathbf{T}$ is a theory of truth just in case, for every ground model M, we have that:

$$\forall \sigma \in Sen(L) : \mathcal{C}_M(\sigma) = \mathbf{a} \Leftrightarrow \mathbf{T}_M(\sigma) = \top_{\mathbf{T}}, \;\; \mathcal{C}_M(\sigma) = \mathbf{d} \Leftrightarrow \mathbf{T}_M(\sigma) = \bot_{\mathbf{T}} \quad (5.8)$$

$$\forall \sigma \in Sen(L_T) : \mathbf{T}_M(T(\overline{\sigma})) = \mathbf{T}_M(\sigma) \quad (5.9)$$

$$\mathbf{T}_M(T(c)) = \bot_{\mathbf{T}} \text{ whenever } I(c) \notin Sen(L_T) \quad (5.10)$$

That is, $\mathbf{T}_M$ should (5.8) *respect the world* and (5.9) *the identity of truth*, while (5.10) *all truth ascriptions to non sentences are valuated as* $\bot_{\mathbf{T}}$. □

We will be particulary interested in theories of truth which output *Strong Kleene valuations*.

Definition 5.2 Strong Kleene valuations
Let $V_M : Sen(L_T) \to \mathbf{V}$ be a valuation of L_T in M such that $\mathbf{V}$ has cardinality 2, 3 or 4. We say that V_M is a Strong Kleene valuation just in case V_M can be described via a lattice $\mathbf{V}_{\leq} = (\mathbf{V}, \leq)$ such that:

- Negation maps the top element of $\mathbf{V}_{\leq}$ to its bottom and vice versa, while it acts as the identity on all elements of $\mathbf{V}_{\leq}$ that are neither top nor bottom.

[5] Modulo our symbolism which reflects that we interpret the semantic values (directly) in assertoric terms.

[6] The range of $\mathbf{T}_M$ may depend on M, i.e., for some M, the range of $\mathbf{T}_M$ may be a strict subset of $\mathbf{V}$.

- Conjunction and disjunction act as meet and join in $\mathbf{V}_{\leq}$.

- Universal and existential quantification behave as generalized conjunction and disjunction respectively. □

Observe that the notion of a Strong Kleene valuation does not mention the semantic behavior of the truth predicate, nor the relation with the valuation of L as induced by the ground model M. It will turn out to be convenient to separate the notion of a Strong Kleene valuation from the notion of *fixed point valuation*, by which we mean a Strong Kleene valuation which respects the defining clauses of a theory of truth.

Definition 5.3 Fixed point valuations and $\mathbf{FP}_M$
Let $V_M : Sen(L_T) \to \mathbf{V}$ be Strong Kleene valuation of L_T in M. We say that V_M is a *fixed point valuation over* M just in case V_M satisfies clauses (5.8), (5.9) and (5.10) of Definition 5.1. We will use $\mathbf{FP}_M$ to denote the set of all 2 and 3 (but not 4!) valued fixed point valuations over M. □

A Strong Kleene theory of truth (*SK theory*) is a theory of truth which assigns a fixed point valuation to each ground model M.

Definition 5.4 SK theory of truth
Let $\mathbf{T}$ be a theory of truth. We say that $\mathbf{T}$ is an *SK theory* just in case, for every ground model M, $\mathbf{T}_M$ is a fixed point valuation. A SK theory which recognizes 3 (4) semantic values is called a SK_3 *theory* (SK_4 *theory*). □

Note that there are no *SK theories* which recognize only two semantic values, as is testified by a ground model which contains a Liar. On the other hand, some ground models M allow for a two valued fixed point valuation of L_T. Also, note that the definition of a SK theory is quite liberal. A "genuine" SK theory $\mathbf{T}$ must, arguably, consist of a *systematic way* in which an arbitrary ground model M is converted into a fixed point valuation $\mathbf{T}_M$, and the notion of a "systematic conversion" does not appear in our definition. However, the definition as given is just fine for our purposes.

Two interesting SK_3 theories are Kripke's *Strong Kleene minimal fixed point theory* $\mathcal{K}$, and his *Strong Kleene maximal intrinsic fixed point theory* $\mathcal{K}^+$. In order to define those theories, we define the following partial order on $\mathbf{FP}_M$. With $V_M, V'_M \in \mathbf{FP}_M$, we let:

$$V_M < V'_M \Leftrightarrow \forall \sigma \in Sen(L_T) : V_M(\sigma) = \mathbf{a} \Rightarrow V'_M(\sigma) = \mathbf{a}$$

When $V_M < V'_M$ we say that V'_M *respects* V_M. The following definitions are all taken from Fitting [16]. We say that V_M is *maximal* just in case for no V'_M we have that $V_M < V'_M$, *minimal* just in case for no V'_M we have that $V'_M < V_M$. We say that V_M and V'_M are *compatible* just in case there exists a fixed point V^*_M which extends them both: $V_M < V^*_M$ and $V'_M < V^*_M$. A fixed point V_M is called *intrinsic* just in case it is compatible with every other fixed point. For any ground model M, we let $\mathbf{I}_M$ be the set of all intrinsic fixed points over M. As Kripke [33] shows, $\mathbf{I}_M$ has a maximum element and $\mathbf{FP}_M$ has a minimal element with respect to the relation $<$. Using the notions just defined, the "official" (Kripkean) definition of $\mathcal{K}$ and $\mathcal{K}^+$ can be given.

Definition 5.5 Kripke's definition of $\mathcal{K}$ and $\mathcal{K}^+$
Let M be a ground model. According to the theory $\mathcal{K}^+$, the valuation of L_T in M is given by $\mathcal{K}^+_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$, where $\mathcal{K}^+_M$ is the maximum of $\mathbf{I}_M$. According to the theory $\mathcal{K}$, the valuation of L_T in M is given by $\mathcal{K}_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$, where $\mathcal{K}_M$ is the minimum of $\mathbf{FP}_M$. □

In order to define the notion of a Generalized Strong Kleene theory of truth precisely, we introduce the notion of a *GSK lattice* as follows. We stipulate that the lattices associated with 3 and 4 valued fixed points are, up to isomorphism, the only *GSK lattices* of cardinality 3, respectively 4 and also, that there do not exist GSK lattices of cardinality 1 or 2. We call the 3 and 4 valued GSK lattices the *GSK base lattices*. When $S_\leq = (S, \leq)$ is a GSK lattice, the lattice $S'_\leq = (S', \leq')$, where $S' = S \cup \{\mathbf{a}_x, \mathbf{d}_x\}$ and where $\leq'$ is the extension of $\leq$ to S which turns $\mathbf{a}_x$ and $\mathbf{d}_x$ into the top and bottom element of S', is also a GSK lattice: we say that $S'_\leq$ is the *direct GSK superlattice* of $S_\leq$. The set of all GSK lattices is the smallest set of lattices which contains the GSK base lattices and which is closed under the formation of direct GSK superlattices. As an immediate consequence of the definition of a GSK lattice, it follows that for each $n \geq 3$ there is, up to isomorphism, only one GSK lattice of cardinality n. Moreover, each GSK lattice of odd cardinality has a linear order, whereas each GSK lattice of even cardinality has a genuine partial order. By deleting the top and bottom element of a GSK lattice $S_\leq$ that is not a base lattice, we obtain another GSK lattice, the *direct GSK* sublattice of $S_\leq$. We say that $S'_\leq$ is a GSK sublattice of $S_\leq$ just in case $S'_\leq = S_\leq$ or $S'_\leq$ can be obtained from $S_\leq$ via a path of direct GSK sublattices. We are now ready to define the notion of a GSK theory.

Definition 5.6 GSK **theory of truth**
Let $\mathbf{T}$ be a theory of truth. We say that $\mathbf{T}$ is a GSK theory just in case in each ground model M, the semantic valuation $\mathbf{T}_M : Sen(L_T) \to \mathbf{V}$ of $\mathbf{T}$ is described via a GSK lattice $\mathbf{V}_\leq = (\mathbf{V}, \leq_\mathbf{V})$ such that, according to $\mathbf{T}_M$:

– Conjunction and disjunction behave as meet and join on $\mathbf{V}_\leq$.
– Universal and existential quantification behave as generalized conjunction and disjunction respectively.
– Negation interchanges the top element with the bottom element of each transitive sublattice of $\mathbf{V}_\leq$ and, also, it acts as the identity on the elements of $\mathbf{V}$ that are neither the top nor bottom element in any GSK sublattice of $\mathbf{V}_\leq$.

When $\mathbf{T}$ is a GSK theory which recognizes n distinct semantic values, we say that $\mathbf{T}$ is a GSK_n *theory*. □

Indeed, any SK theory is a GSK theory. Before we state the *assertoric transfer theorem*, which will give us a recipe for defining GSK theories, we first define the notion of one GSK theory respecting another in the expected manner.

Definition 5.7 $\mathbf{T}'$ respects $\mathbf{T}$.
Let $\mathbf{T}$ and $\mathbf{T}'$ be two GSK theories, having $\top$ and $\top'$ as their respective top values. We say that $\mathbf{T}'$ *respects* $\mathbf{T}$ just in case, for every ground model M and $\sigma \in Sen(L_T)$, it holds that:

$$\mathbf{T}_M(\sigma) = \top \Rightarrow \mathbf{T}'_M(\sigma) = \top'$$

When $\mathbf{T}'$ respects $\mathbf{T}$ we write $\mathbf{T} < \mathbf{T}'$. □

Theorem 5.1 Assertoric transfer theorem
Let $\mathbf{T}^3$ be a SK_3 theory and let $\diamond$ denote the semantic value that is neither top nor bottom in the lattice of $\mathbf{T}^3$. Let $\mathbf{T}^n$ be a GSK_n theory over a lattice $\mathbf{V}^n_{\leq}$ whose top and bottom element are denoted by $\top$ and $\bot$ respectively. For each ground model M, define $\mathbf{T}^{n+2}_M$ as follows:

$$\mathbf{T}^{n+2}_M(\sigma) = \begin{cases} \mathbf{x}^*, & \mathbf{T}^n_M(\sigma) = \mathbf{x} \text{ and } \mathbf{T}^3_M(\sigma) \neq \diamond \\ \mathbf{x}, & \mathbf{T}^n_M(\sigma) = \mathbf{x} \text{ and } \mathbf{T}^3_M(\sigma) = \diamond \end{cases} \tag{5.11}$$

If $\mathbf{T}^3 < \mathbf{T}^n$, then $\mathbf{T}^{n+2}$ is a GSK_{n+2} theory over the lattice $\mathbf{V}^{n+2}_{\leq}$, which is obtained as the extension of the lattice $\mathbf{V}^n_{\leq}$ by adding $\top^*$ and $\bot^*$ as (new) top and bottom element respectively.
Proof: From the fact that both $\mathbf{T}^n$ and $\mathbf{T}^3$ are GSK theories, that $\mathbf{T}^n$ respects $\mathbf{T}^3$ and the definition of $\mathbf{T}^{n+2}$. □

5.4 The Method of Closure Games

By an *AD sentence*, we mean a signed, with A(ssertible) or D(eniable), sentence of L_T. $\mathcal{X}$ denotes the set of all AD sentences.

$$\mathcal{X} = \{X_\sigma \mid X \in \{A, D\}, \sigma \in Sen(L_T)\}$$

With $At(L)$, we denote the set of atomic sentences of L. These sentences are assumed to receive their (classical) valuation from the ground model M and can be thought of as the "non-semantic facts". We will treat (atomic) truth ascriptions to non-sentential objects on a par with members of $At(L)$. Hence, it is convenient to define, with $M = (D, I)$, the set $At^*_M(L)$ as follows:

$$At^*_M(L) = At(L) \cup \{T(c) \mid I(c) \notin Sen(L_T)\}$$

We assume, in line with Definition 5.1, that (atomic) sentences which ascribe truth to non-sentential objects, always have to be denied. This assumption leads to the following definition of the *world* w_M associated with ground model M:

$$w_M = \{A_\sigma \mid \mathcal{C}_M(\sigma) = \mathbf{a}, \sigma \in At(L)\} \cup \{D_\sigma \mid \mathcal{C}_M(\sigma) = \mathbf{d}, \sigma \in At(L)\} \cup \{D_{T(c)} \mid I(c) \notin Sen(L_T)\}$$

An AD sentence X_σ is either of conjunctive type $\sqcap$ or of disjunctive type $\sqcup$ and has a set of *immediate AD subsentences*, $\Pi(X_\sigma)$. We depict the information just mentioned in the form of an *assertoric rule*:

$$\frac{X_\sigma}{\Pi(X_\sigma)}\sqcup \qquad \frac{X_\sigma}{\Pi(X_\sigma)}\sqcap$$

The *assertoric rules*[7] are stated in the table below. As testified by, amongst others, the rules for the truth predicate T, the assertoric rules depend on the

[7] In the rules for T, $\overline{\sigma} \in Con(L_T)$ is a quotational or non quotational constant which denotes σ in M. In the rules for the quantifiers, $\phi(x/t)$ denotes the result of the uniform replacement of variable x by constant t in $\phi(x)$. For (signed) negations, truth ascriptions and atomic sentences of L, it does not matter which type, $\sqcup$ or $\sqcap$, they are given. The actual allotment of types to those sentences as displayed below was chosen for sake of symmetry only.

details of sentential reference and are, accordingly, defined relative to a ground model M.

$\neg$	$\dfrac{A_{\neg\alpha}}{\{D_\alpha\}}\sqcap$	$\dfrac{D_{\neg\alpha}}{\{A_\alpha\}}\sqcup$
$\vee$	$\dfrac{A_{(\alpha\vee\beta)}}{\{A_\alpha, A_\beta\}}\sqcup$	$\dfrac{D_{(\alpha\vee\beta)}}{\{D_\alpha, D_\beta\}}\sqcap$
$\wedge$	$\dfrac{A_{(\alpha\wedge\beta)}}{\{A_\alpha, A_\beta\}}\sqcap$	$\dfrac{D_{(\alpha\wedge\beta)}}{\{D_\alpha, D_\beta\}}\sqcup$
$\exists$	$\dfrac{A_{\exists x\phi(x)}}{\{A_{\phi(x/t)} \mid t \in Con(L_T)\}}\sqcup$	$\dfrac{D_{\exists x\phi(x)}}{\{D_{\phi(x/t)} \mid t \in Con(L_T)\}}\sqcap$
$\forall$	$\dfrac{A_{\forall x\phi(x)}}{\{A_{\phi(x/t)} \mid t \in Con(L_T)\}}\sqcap$	$\dfrac{D_{\forall x\phi(x)}}{\{D_{\phi(x/t)} \mid t \in Con(L_T)\}}\sqcup$
T	$\dfrac{A_{T(\overline{\sigma})}}{\{A_\sigma\}}\sqcap$	$\dfrac{D_{T(\overline{\sigma})}}{\{D_\sigma\}}\sqcup$
$\sigma \in At^\star_M(L)$	$\dfrac{A_\sigma}{\{A_\sigma\}}\sqcap$	$\dfrac{D_\sigma}{\{D_\sigma\}}\sqcup$

Figure 5.2: The assertoric rules of L_T

The notions of a *strategy*, an *expansion*, a *closure condition* and a *valuation induced by a closure condition* were discussed in the introduction. The definition below summarizes this discussion.

Definition 5.8 Strategies, expansions, closure conditions, valuations
1. A *strategy for player* $\sqcup$ is a function f which maps each X_σ of type $\sqcup$ to one element of $\Pi(X_\sigma)$. The set of all strategies of player $\sqcup$ is denoted by $\mathcal{F}$.
2. A *strategy for player* $\sqcap$ is a function g which maps each X_σ of type $\sqcap$ to one element of $\Pi(X_\sigma)$. The set of all strategies of player $\sqcap$ is denoted by $\mathcal{G}$.
3. With $f \in \mathcal{F}$, $g \in \mathcal{G}$ and $X_\sigma \in \mathcal{X}$, $\mathsf{exp}(X_\sigma, f, g)$ denotes *the expansion of* X_σ *by* f *and* g. The set of all expansions in M is[8] denoted by EXP_M.
4. A *closure condition* $\dagger(M) = \{O^\dagger_M, C^\dagger_M\}$ is a bipartition of EXP_M into the sets $O^\dagger_M \neq \varnothing$, consisting of the *open*$_\dagger$ *expansions in* M, and $C^\dagger_M \neq \varnothing$, containing the *closed*$_\dagger$ *expansions*[9] in M.
5. A closure condition $\dagger(M) = \{O^\dagger_M, C^\dagger_M\}$ gives rise to closure conditions for AD sentences:

$$O^\dagger_M(X_\sigma) \Leftrightarrow \exists f \in \mathcal{F} \forall g \in \mathcal{G} : \mathsf{exp}(X_\sigma, f, g) \in O^\dagger_M$$

$$C^\dagger_M(X_\sigma) \Leftrightarrow \text{ not } O^\dagger_M(X_\sigma)$$

6. The closure conditions for AD sentences are used to induce $\mathcal{V}^\dagger_M$:
$\mathcal{V}^\dagger_M(\sigma) = \mathbf{a} \Leftrightarrow O^\dagger_M(A_\sigma)$ & $C^\dagger_M(D_\sigma)$ $\quad$ $\mathcal{V}^\dagger_M(\sigma) = \mathbf{b} \Leftrightarrow O^\dagger_M(A_\sigma)$ & $O^\dagger_M(D_\sigma)$
$\mathcal{V}^\dagger_M(\sigma) = \mathbf{d} \Leftrightarrow C^\dagger_M(A_\sigma)$ & $O^\dagger_M(D^\sigma)$ $\quad$ $\mathcal{V}^\dagger_M(\sigma) = \mathbf{n} \Leftrightarrow C^\dagger_M(A_\sigma)$ & $C^\dagger_M(D_\sigma)$ $\quad \square$

[8]The assertoric rules for truth testify that the set of all expansions depends on the ground model under consideration.

[9]The condition that $C^\dagger_M$ and $O^\dagger_M$ are non empty rules ensures that we do not have to consider the possibility that $\mathcal{V}^\dagger_M$ valuates all L_T sentences as $\mathbf{n}$ ($O^\dagger_M = \varnothing$) or as $\mathbf{b}$ ($C^\dagger_M = \varnothing$), ensuring that $\mathcal{V}^\dagger_M$ is at least 2 valued. This feature will be convenient for the formulation of theorems that follow.

Here are some notational conventions.

Definition 5.9 Some notational conventions
In this paper, the non-quotational constants λ, τ, η, θ and μ will be used as follows, where I is some interpretation function.

1. $I(\lambda) = \neg T(\lambda)$. We say that $\neg T(\lambda)$ is a *Liar*.
2. $I(\tau) = T(\tau)$. We say that $T(\tau)$ is a *Truthteller*.
3. $I(\eta) = T(\eta) \vee \neg T(\eta)$. We say that $T(\eta) \vee \neg T(\eta)$ is a *Tautologyteller*.
4. $I(\theta) = T(\theta) \wedge \neg T(\theta)$. We say that $T(\theta) \wedge \neg T(\theta)$ is a *Contradictionteller*.
5. $I(\mu) = T(c_0)$ where, for each n, $I(c_n) = \neg T(c_{n+1})$. We say that $T(\mu)$ is an *Unstabilityteller*.

To be sure, the notational convention does not imply that every ground model contains one of the five sentences just defined. However, if we use a sentence which is build with the constant λ, τ, η, θ or μ, we always presuppose a ground model in which a Liar, Truthteller, Tautologyteller, Contradictionteller or Unstabilityteller occurs. □

Here are six examples of expansions, which will be used to illustrate some convenient classifications of expansions.

1. $A_{P(c)\vee T(\lambda)}, A_{P(c)}, A_{P(c)}, \ldots$, where $A_{P(c)} \in w_M$
2. $A_{\neg P(c)}, D_{P(c)}, D_{P(c)}, \ldots$ where $D_{P(c)} \notin w_M$
3. $A_{P(c)\vee T(\tau)}, A_{T(\tau)}, A_{T(\tau)}, A_{T(\tau)}, \ldots$
4. $D_{T(\tau)}, D_{T(\tau)}, D_{T(\tau)}, \ldots$
5. $A_{T(\mu)}, A_{T(c_0)}, A_{\neg T(c_1)}, D_{T(c_1)}, D_{\neg T(c_2)}, A_{T(c_2)} \ldots$
6. $A_{T(\lambda)}, A_{\neg T(\lambda)}, D_{T(\lambda)}, D_{\neg T(\lambda)}, A_{T(\lambda)}, \ldots$

First, observe that every expansion is either *stable*$_A$, *stable*$_D$ or *unstable*. The formal definition of these notions is clear from the remark that expansions 1 and 3 are stable$_A$, 2 and 4 are stable$_D$ and 5 and 6 are unstable. Next, observe that every expansion is either *grounded* or *ungrounded*, where an expansion is grounded just in case it contains, for some $\sigma \in At^*_M(L)$, X_σ; we say that X_σ is the *ground* of the expansion. Grounded expansions are either *correct in* M or *incorrect in* M. An expansion is correct in M just in case its ground is contained in w_M, incorrect if its ground is not contained in w_M. Thus, expansion 1 is a grounded and correct expansion, while expansion 2 is grounded and incorrect. An (ungrounded) expansion is *vicious* just in case it contains a *vicious cycle*, or in other words, an expansion $\{y_n\}_{n\in\mathbb{N}}$ is vicious just in case:

$$\exists\sigma\forall n\exists m, m' > n : y_m = A_\sigma \text{ and } y_{m'} = D_\sigma$$

Indeed, expansion 6 is vicious. We introduce the following abbreviations for subsets of EXP_M.

Definition 5.10 Classifying expansions
We define the following subsets of EXP_M.

- G_M: the set of all grounded expansions.
  U_M: the set of all ungrounded expansions.

- G_M^{cor}: the set of all grounded and correct expansions.
  G_M^{inc}: the set of all grounded and incorrect expansions.

- U_M^{vic}: the set of all (ungrounded) vicious expansions.
  U_M^{nvi}: the set of all ungrounded non-vicious expansions.

- US_M^A: the set of all ungrounded stable$_A$ expansions.
  US_M^D: the set of all ungrounded stable$_D$ expansions.
  UU_M: the set of all (ungrounded) unstable expansions. □

For any expansion exp, we let exp' denote the *successor expansion* of exp, by which we mean the expansion that is obtained by removing the first term of exp. A closure condition $\dagger(M) = \{O_M^\dagger, C_M^\dagger\}$ satisfies the *stable judgement constraint* (SJC), just in case, for every expansion $\mathsf{exp} \in \mathsf{EXP_M}$, it holds that:

$$\mathsf{SJC}: \mathsf{exp} \in C_M^\dagger \Leftrightarrow \mathsf{exp}' \in C_M^\dagger$$

Note that, equivalently, SJC can be formulated in terms of openness:

$$\mathsf{SJC}: \mathsf{exp} \in O_M^\dagger \Leftrightarrow \mathsf{exp}' \in O_M^\dagger$$

The SJC will be the central notion of our two stable judgement theorems. Below, we prove the first stable judgement theorem, in which we refer to the set of all *AD* subsentences of X_σ, denoted $\overline{\Pi}(X_\sigma)$. Formally, $\overline{\Pi}(X_\sigma)$ is defined by taking the transitive closure of the binary relation induced by the set of all *immediate* *AD* subsentences of X_σ:

- $\Pi(\cdot,\cdot)$ is defined by: $\Pi(X_\sigma, Y_\alpha) \Leftrightarrow Y_\alpha \in \Pi(X_\sigma)$.

- $\overline{\Pi}(\cdot,\cdot)$ is defined as the transitive closure of $\Pi(\cdot,\cdot)$.

- $\overline{\Pi}(\cdot)$ is defined by: $\overline{\Pi}(X_\sigma) = \{Y_\alpha \mid \overline{\Pi}(X_\sigma, Y_\alpha)\}$

Theorem 5.2 First stable judgement theorem
Let $M = (D, I)$ be a ground model, let $\dagger(M) = \{O_M^\dagger, C_M^\dagger\}$ be a closure condition which satisfies SJC and let $\mathcal{V}_M^\dagger$ be the valuation function induced by $\dagger(M)$. It holds that:

1. $\mathcal{V}_M^\dagger : Sen(L_T) \to \{\mathbf{a}, \mathbf{b}, \mathbf{n}, \mathbf{d}\}$ is either a 2-, 3- or 4-valued Strong Kleene valuation (see Definition 5.2) which respects one of the lattices indicated in Figure 5.3.

2. For each $\sigma \in Sen(L_T)$ it holds that $\mathcal{V}_M^\dagger(T(\overline{\sigma})) = \mathcal{V}_M^\dagger(\sigma)$. That is, if $\dagger(M)$ satisfies SJC then $\mathcal{V}_M^\dagger$ respects the identity of truth.

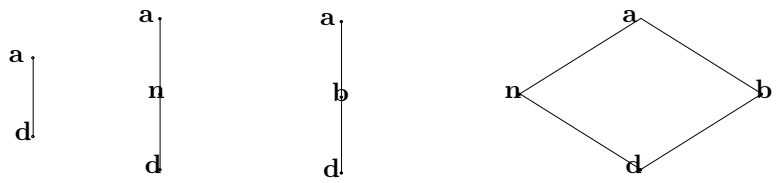


Figure 5.3: Three linear orders, one partial order.

Proof: Let $\dagger(M) = \{O^{\dagger}_M, C^{\dagger}_M\}$ be a closure condition which satisfies SJC. Notice that, in order to show that $\mathcal{V}^{\dagger}_M$ is Strong Kleene valuation which respects the identity of truth, it suffices to show that for every AD sentence X_σ:

$$\text{type of } X_\sigma = \sqcup \Rightarrow \left(O^{\dagger}_M(X_\sigma) \Leftrightarrow \exists Y_\alpha \in \Pi(X_\sigma) : O^{\dagger}_M(Y_\alpha)\right)$$

$$\text{type of } X_\sigma = \sqcap \Rightarrow \left(O^{\dagger}_M(X_\sigma) \Leftrightarrow \forall Y_\alpha \in \Pi(X_\sigma) : O^{\dagger}_M(Y_\alpha)\right)$$

We illustrate for $A_{\alpha\wedge\beta}$. Other cases are similar and left to the reader.
$\Rightarrow$ Suppose that $O^{\dagger}_M(A_{\alpha\wedge\beta})$. This means that there is a strategy $f \in \mathcal{F}$ such that for all $g \in \mathcal{G}$, $\mathsf{exp}(A_{\alpha\wedge\beta}, f, g)$ is open$_\dagger$. Now $A_{\alpha\wedge\beta}$ is of type $\sqcap$, and the strategies of player $\sqcap$ can be bi-partitioned into strategies g_α, which have $g(A_{\alpha\wedge\beta}) = A_\alpha$ and strategies of type g_β, which have $g(A_{\alpha\wedge\beta}) = A_\beta$. As f results in an open expansion, no matter whether player $\sqcap$ plays a strategy of type g_α or g_β, it follows, as $\dagger(M)$ satisfies SJC, that f is such that for all $g \in \mathcal{G}$, we have that $\mathsf{exp}(A_\alpha, f, g) \in O^{\dagger}_M$ and that $\mathsf{exp}(A_\beta, f, g) \in O^{\dagger}_M$. Hence, $O^{\dagger}_M(A_\alpha)$ and $O^{\dagger}_M(A_\beta)$.
$\Leftarrow$ Suppose that $O^{\dagger}_M(A_\alpha)$ and $O^{\dagger}_M(A_\beta)$. This means that there exists a strategy $f_\alpha \in \mathcal{F}$ such that for all $g \in \mathcal{G}$ we have that $\mathsf{exp}(A_\alpha, f_\alpha, g) O^{\dagger}_M$ and that there exists a strategy $f_\beta \in \mathcal{F}$ such that for all $g \in \mathcal{G}$ we have that $\mathsf{exp}(A_\beta, f_\beta, g) \in O^{\dagger}_M$. Let $f \in \mathcal{F}$ be any strategy which satisfies:

- $X_\sigma \in \overline{\Pi}(A_\alpha), \text{type of } X_\sigma = \sqcup \Rightarrow f(X_\sigma) = f_\alpha(X_\sigma)$
- $X_\sigma \in (\overline{\Pi}(A_\beta) - \overline{\Pi}(A_\alpha)), \text{type of } X_\sigma = \sqcup \Rightarrow f(X_\sigma) = f_\beta(X_\sigma)$

From the fact that $\dagger(M)$ satisfies SJC, it follows that the constructed f is such that for all $g \in \mathcal{G}$ we have that $\mathsf{exp}(A_{\alpha\wedge\beta}, f, g) \in O^{\dagger}_M$. □

Thus, picking a closure condition which satisfies SJC ensures that we induce a Strong Kleene valuation which respects the identity of truth. As such, a closure condition which satisfies SJC does not guarantee that we induce a fixed point valuation (as defined by Definition 5.3). However, by posing the following additional constraint on closure conditions, we ensure that they induce fixed point valuations. Let $\dagger(M) = \{O^{\dagger}_M, C^{\dagger}_M\}$ be any closure condition. We say that $\dagger(M)$ satisfies the *world respecting constraint*, WRC, just in case:

$$\mathsf{WRC} : G^{cor}_M \subseteq O^{\dagger}_M \text{ and } G^{inc}_M \subseteq C^{\dagger}_M$$

We get the following corollary to Theorem 5.2.

Corollary 5.1 Inducing fixed point valuations
Let $\dagger(M) = \{O^{\dagger}_M, C^{\dagger}_M\}$ be a closure condition which satisfies WRC and SJC.

Then $\mathcal{V}^{\dagger}_{M}$ is a (2-, 3- or 4-valued) fixed point valuation in the sense of Definition 5.3.
Proof: In light of Theorem 5.2, it suffices to show that if $\dagger(M)$ satisfies WRC, then $\mathcal{V}^{\dagger}_{M}$ satisfies clauses (5.8) and (5.10) of Definition 5.1. A proof can be given by induction on the complexity of sentences of L, accounting for the non-sentential truth ascriptions in a straightforward way. This is left to the reader. □

Let us put the first stable judgement theorem to work. Consider the following closure conditions.

gr(oundedness) closure conditions: $O^{gr}_{M} = G^{cor}_{M}$

⧫ closure conditions: $O^{\blacklozenge}_{M} = G^{cor}_{M} \cup U^{nvi}_{M}$

It is easily seen that those closure conditions satisfy SJC and WRC. Hence, by the first stable judgement theorem, $\mathcal{V}^{gr}$ and $\mathcal{V}^{\blacklozenge}$ are SK theories of truth. In fact, we have that:

Proposition 5.1 $\mathcal{V}^{gr} = \mathcal{K}$ **whereas** $\mathcal{V}^{\blacklozenge}$ **is a** SK_4 **theory.**
Proof: In Wintein [62] we showed that the groundedness closure conditions induce Kripke's Strong Kleene minimal fixed point theory of truth, i.e., that $\mathcal{V}^{gr} = \mathcal{K}$. For sake of completeness, the proof is also given in Appendix I of this chapter. The fact that $\mathcal{V}^{\blacklozenge}$ is an SK_4 theory follows from the observation (that the ⧫ closure conditions satisfy SJC and WRC and) that $\mathcal{V}^{\blacklozenge}_{M}(\neg T(\lambda)) = \mathbf{n}$ while $\mathcal{V}^{\blacklozenge}_{M}(T(\tau)) = \mathbf{b}$. □

It is instructive to explain, in terms of MCG, why $\mathcal{V}^{gr}_{M}$ is 3-valued, whereas $\mathcal{V}^{\blacklozenge}_{M}$ is 4-valued[10]. Here we go. Let M be a ground model and let X^{-1} denote the *AD inverse* of X: $A^{-1} = D$, $D^{-1} = A$. With $X \in \{A, D\}$, we have that:

$$O^{gr}_{M}(X_\sigma) \Rightarrow C^{gr}_{M}(X^{-1}_\sigma) \tag{5.12}$$

Equation (5.12) follows from (5.13):

$$\exists f \forall g\ \mathsf{exp}(X_\sigma, f, g) \in O^{gr}_{M} \Rightarrow \exists g \forall f\ \mathsf{exp}(X^{-1}_\sigma, f, g) \in C^{gr}_{M} \tag{5.13}$$

Before we establish (5.13), we first point out that (5.12) follows from (5.13). Suppose that $O^{gr}_{M}(X_\sigma)$. Thus, player ⊔ can ensure that an expansion which start with X_σ ends up in O^{gr}_{M}. This implies, via equation (5.13), that player ⊓ can ensure that an expansion which start with X^{-1}_σ ends up in C^{gr}_{M}. But the latter means, as O^{gr}_{M} and C^{gr}_{M} bipartition EXP_M, that player ⊔ cannot ensure than an expansion which starts with X^{-1}_σ ends in O^{gr}_{M}. Accordingly, $C^{gr}_{M}(X^{-1}_\sigma)$.

Equation (5.13) follows from a general observation. To state that observation, we define, for any expansion $\mathsf{exp} = \{y_n\}_{n\in\mathbb{N}}$, its *inverse expansion* $\mathsf{exp}^{-1} = \{z_n\}_{n\in\mathbb{N}}$ by letting, for any $n \in \mathbb{N}$:

$$z_n = A_\sigma \Leftrightarrow y_n = D_\sigma$$

[10] In fact, one can show that $\mathcal{V}^{gr}_{M}$ is 3 valued for every ground model M, whereas $\mathcal{V}^{\blacklozenge}_{M}$ is, depending on the ground model, either 3 or 4 valued.

Further, for any set $S \subseteq \mathsf{EXP}_M$ of expansions, we define its inverse S^{-1} by letting $S^{-1} = \{\mathsf{exp}^{-1} \mid \mathsf{exp} \in S\}$. For each strategy f of player $\sqcup$, there is a *mirror strategy* for player $\sqcap$, call it g_f, which is defined as follows:

$$g_f(X_\alpha) = Y_\beta \Leftrightarrow f(X_\alpha^{-1}) = Y_\beta^{-1}$$

Similarly, for each strategy g of player $\sqcap$, there is a mirror strategy for player $\sqcup$ which may be called f_g. Let S be any set of expansions. From an inspection of the notion of a mirror strategy, it follows that:

$$\exists f \forall g\ \mathsf{exp}(X_\sigma, f, g) \in S \Leftrightarrow \exists g \forall f\ \mathsf{exp}(X_\sigma^{-1}, f, g) \in S^{-1} \tag{5.14}$$

As the set of expansions G_M^{cor} is the inverse of G_M^{inc}, it follows from (5.14) that:

$$\exists f \forall g\ \mathsf{exp}(X_\sigma, f, g) \in G_M^{cor} \Leftrightarrow \exists g \forall f\ \mathsf{exp}(X_\sigma^{-1}, f, g) \in G_M^{inc} \tag{5.15}$$

From (5.15) we get (5.13) and, accordingly (5.12). The principle that is underlying the 3 valuedness of $\mathcal{V}^{gr}$, i.e., (5.12), breaks down for $\mathcal{V}^{\blacklozenge}$. For, we have that:

$$O_M^{\blacklozenge}(X_\sigma) \not\Rightarrow C_M^{\blacklozenge}(X_\sigma^{-1}) \tag{5.16}$$

The reason for this is that the set of expansions U_M^{nvi}, which is open$_{\blacklozenge}$, is its own inverse. Hence, the fact that player $\sqcup$ can force an expansion of A_σ to end up in U_M^{nvi} implies that player $\sqcap$ can force the expansion of D_σ to end up in $(U_M^{nvi})^{-1} = U_M^{nvi}$. But the fact that player $\sqcap$ can force the expansion of D_σ to end up in U_M^{nvi} does not preclude the possibility that player $\sqcup$ may as well be able to force D_σ to end up in U_M^{nvi}. Hence, A_σ and D_σ may both be open$_{\blacklozenge}$. The previous remarks are illustrated by considering the two expansions of the Truthteller:

$$A_{T(\tau)}, A_{T(\tau)}, A_{T(\tau)}, \ldots \qquad\qquad D_{T(\tau)}, D_{T(\tau)}, D_{T(\tau)}, \ldots$$

Before we state our second stable judgement theorem, it is instructive to compare the $\blacklozenge$ closure conditions with the $\Diamond$ closure conditions, that are defined below. Before we define the the $\Diamond$ closure conditions, observe that the $\blacklozenge$ closure conditions allow for the following, equivalent, definition:

$$\blacklozenge \textbf{ closure conditions}\text{: } C_M^{\blacklozenge} = G_M^{inc} \cup U_M^{vic}$$

This reformulation is convenient as it clearly lays bare the distinction with the $\Diamond$ closure conditions:

$$\Diamond \textbf{ closure conditions}\text{: } C_M^{\Diamond} = G_M^{inc} \cup \{\mathsf{exp} \mid \exists \sigma \in Sen(L_T) : A_\sigma,\ D_\sigma \text{ on } \mathsf{exp}\ \}$$

So the only difference between the $\blacklozenge$ closure conditions and the $\Diamond$ closure conditions is that the former considers all expansions closed which contain A_σ and D_σ *in a cycle*, whereas the latter does away with the condition of cyclicality: whenever an expansion contains an "AD clash" it is closed, whether or not this clash occurs in a cycle . To illustrate the difference between the $\blacklozenge$ and the $\Diamond$ closure conditions, we consider the following expansion of a denial of the Tautologyteller:

$$D_{T(\eta)\vee\neg T(\eta)}, D_{\neg T(\eta)}, A_{T(\eta)}, A_{T(\eta)\vee\neg T(\eta)}, A_{T(\eta)}, A_{T(\eta)\vee\neg T(\eta)}, \ldots \tag{5.17}$$

This expansion is open according to the ⧫ closure conditions—as the "AD clash" does not occur in a cycle—while it is closed according to the ◊ closure conditions. The successor expansion of (5.17), however, is open according to the ◊ closure conditions, which establishes that these closure conditions do not satisfy SJC.

$\mathcal{V}^{\Diamond}$ defines a 4-valued theory of truth of which it can be shown[11] that it does not have a compositional semantics. This raises the question whether satisfying SJC is, besides a sufficient condition, also a necessary condition for closure conditions to induce a SK valuation which respects the identity of truth. The answer to that question is 'no', as testified by the following proposition.

Proposition 5.2 SJC **and** SK **compositionality come apart**.
Proof: The ⋆ closure conditions, stated below, violate SJC while they define a 3-valued SK theory of truth. In the definition of the ⋆ closure conditions, c is an arbitrary non-quotational constant of L_T.

$$O^{\star}_M = G^{cor}_M \cup \{\mathsf{exp} \mid A_{T(c)\vee\neg T(c)} \text{ or } D_{T(c)} \text{ on } \mathsf{exp} \text{ and } I(c) = T(c)\} \qquad (5.18)$$

The ⋆ closure conditions are a (minimal) modification of the (gr)oundedness closure conditions. According to the ⋆ closure conditions, the expansions in G^{cor}_M are open and, besides those, all (and only) the expansions which contain $A_{T(c)\vee\neg T(c)}$ or $D_{T(c)}$ for some c such that $I(c) = T(c)$ are open. A little reflection shows that this ensures that $\mathcal{V}^{\star}_M$ is just like $\mathcal{V}^{gr}_M$, apart from a valuation of Truthtellers—i.e., sentences of form $T(c)$ such that $I(c) = T(c)$—and compounds of Truthtellers. In particular, with $T(\tau)$ a Truthteller, we have that:

$$\mathcal{V}^{\star}_M(T(\tau)) = \mathbf{d}, \qquad \mathcal{V}^{\star}_M(T(\tau) \vee \neg T(\tau)) = \mathbf{a}$$

Being a minimal modification of $\mathcal{V}^{gr}_M$, $\mathcal{V}^{\star}_M$ is a SK_3 theory. However, the ⋆ closure conditions violate SJC, which is easily seen by inspecting the following expansion:

$$A_{T(\tau)\vee\neg T(\tau)}, A_{T(\tau)}, A_{T(\tau)}, A_{T(\tau)} \dots$$

Indeed, this expansion is open according to ⋆ closure conditions as it contains $A_{T(\tau)\vee\neg T(\tau)}$ and as $I(\tau) = T(\tau)$. Its successor expansion, which does not contain $A_{T(\tau)\vee\neg T(\tau)}$ or $D_{T(\tau)}$ is closed, and so the ⋆ closure conditions violate SJC while they induce a SK_3 theory. □

Thus, Proposition 5.2 testifies that the first stable judgement theorem cannot be read in the converse direction. However, the second stable judgement theorem comes close to a converse reading of the first stable judgement theorem: it states that any (2, 3 or 4 valued) SK valuation which respects the identity of truth can be induced from a closure condition which satisfies SJC. Before we state this theorem, we define the notion of the *correctness* of an AD sentence with respect to a (2-, 3- or 4- valued) valuation[12] V_M.

Definition 5.11 V_M **correctness**
Let V_M be a (2-, 3- or 4-valued) valuation for L_T whose range $\mathbf{V}$ is such that

[11] The reader may verify this by considering the sentence $I(c) = \neg T(c) \vee T(\tau)$, where $T(\tau)$ is the Truthteller.
[12] Note: V_M does not need to be Strong Kleene.

$\{\mathbf{a},\mathbf{d}\} \subseteq \mathbf{V} \subseteq \{\mathbf{a},\mathbf{b},\mathbf{n},\mathbf{d}\}$. The notion of V_M *correctness*, applicable to AD sentences, is defined as follows.

$$X_\sigma \text{ is } V_M \text{ correct } \Leftrightarrow (X = A, V_M(\sigma) \in \{\mathbf{a},\mathbf{b}\}) \text{ or } (X = D, V_M(\sigma) \in \{\mathbf{d},\mathbf{b}\})$$

Intuitively, an AD sentence X_σ is V_M correct iff its judgement (Assertible or Deniable) with respect to σ is correct from the standpoint of V_M. □

Theorem 5.3 Second stable judgement theorem
Let M be a ground model and let V_M be a 2-, 3- or 4-valued Strong Kleene valuation for L_T which respects the identity of truth. Then there is a closure condition $\dagger(M)$ which satisfies SJC and which is such that $\mathcal{V}^\dagger_M = V_M$.
Proof:
Let V_M be a 2-, 3- or 4-valued Strong Kleene valuation for L_T which respects the identity of truth. Using the notion of V_M correctness, we define a closure condition $\dagger(M) = \{O^\dagger_M, C^\dagger_M\}$ for which we will show that it satisfies SJC and that it is such that $\mathcal{V}^\dagger_M = V_M$. Let $\mathsf{exp} = \{y_n\}_{n\in\mathbb{N}}$ be an arbitrary expansion of EXP_M. We let:

$$\mathsf{exp} \in O^\dagger_M \Leftrightarrow \exists n \forall m > n : y_m \text{ is } V_M \text{ correct} \tag{5.19}$$

It is clear, from the "limit behavior definition" of $\dagger(M)$, that $\dagger(M)$, satisfies SJC. Note that, in order to show that $\mathcal{V}^\dagger_M = V_M$, it suffices to show that:

$$X_\sigma \text{ is } V_M \text{ correct} \Leftrightarrow \exists f \in \mathcal{F} \forall g \in \mathcal{G} : \mathsf{exp}(X_\sigma, f, g) \in O^\dagger_M$$

$\Rightarrow$ Suppose that X_σ is V_M correct. Observe that, from the fact that V_M is SK and respects the identity of truth, we have:

$$\text{type of } X_\sigma = \sqcup \Rightarrow \exists Y_\alpha \in \Pi(X_\sigma) : Y_\alpha \text{ is } V_M \text{ correct}$$
$$\text{type of } X_\sigma = \sqcap \Rightarrow \forall Y_\alpha \in \Pi(X_\sigma) : Y_\alpha \text{ is } V_M \text{ correct}$$

From these two observations, it readily follows that if we start from an X_σ which is V_M correct, player $\sqcup$ has a strategy, say f, which ensures that, for every $g \in \mathcal{G}$, *all* the terms of $\mathsf{exp}(X_\sigma, f, g)$ are V_M correct. Hence, player $\sqcup$ can ensure that the expansion of X_σ ends up in $O^\dagger_M$.
$\Leftarrow$ Suppose that X_σ is V_M incorrect. Observe that, from the fact that V_M is SK and respects the identity of truth, we have:

$$\text{type of } X_\sigma = \sqcup \Rightarrow \forall Y_\alpha \in \Pi(X_\sigma) : Y_\alpha \text{ is } V_M \text{ incorrect}$$
$$\text{type of } X_\sigma = \sqcap \Rightarrow \exists Y_\alpha \in \Pi(X_\sigma) : Y_\alpha \text{ is } V_M \text{ incorrect}$$

From these two observations, it readily follows that if we start from an X_σ which is V_M incorrect, player $\sqcap$ has a strategy, say g, which ensures that, for every $f \in \mathcal{F}$, *all* the terms of $\mathsf{exp}(X_\sigma, f, g)$ are V_M incorrect. Hence, player $\sqcap$ can ensure that the expansion of X_σ ends up in $C^\dagger_M$, from which it follows that player $\sqcup$ cannot ensure that the expansion of X_σ ends up in $O^\dagger_M$. □

So, in order to induce, say, Kripke's SK maximal intrinsic fixed point $\mathcal{K}^+$, via the MCG, we may define closure conditions, via (5.19), in terms of $\mathcal{K}^+_M$ correctness. Closure conditions for $\mathcal{K}^+$ that are defined as such are parasitic on

Kripke's framework for truth in a way that the gr(oundedness), ⧫ and ◊ closure conditions are not. As $\mathcal{K}^+$ is an interesting theory of truth, it would be bad news for MCG, as a framework for truth, if it had to rely, for the definition of $\mathcal{K}^+$, on notions that are borrowed from an alternative framework. Luckily, the MCG does have access to $\mathcal{K}^+$ via notions that are not borrowed from an alternative framework. In the next section, we see how this works out.

5.5 Assertoric branches and trees

5.5.1 Inducing theory $\mathcal{V}^\bullet$ via branch closure conditions

For any expansion exp, [exp] will denote the set of terms of exp. For any AD sentence X_σ and strategy f of player ⊔, $B_f(X_\sigma)$ denotes the set of terms that occur on some expansion of X_σ relative to f. $B_f(X_\sigma)$ is called the *branch* of X_σ induced by f. To be sure, $B_f(X_\sigma)$ is defined as follows:

$$B_f(X_\sigma) = \bigcup_{g \in \mathcal{G}} [\mathsf{exp}(X_\sigma, f, g)]$$

We will use Branch_M to denote[13]the set of all branches relative to ground model M. The (assertoric) tree of X_σ, $\mathfrak{T}^\sigma_X$, is the set of all its branches. That is:

$$\mathfrak{T}^\sigma_X = \{B_f(X_\sigma) \mid f \in \mathcal{F}\}$$

Branches are judged to be open or closed relative to closure conditions which are applicable to branches; a branch closure condition $\ddagger(M) = \{O^\ddagger_M, C^\ddagger_M\}$ is a bipartition of Branch_M. An assertoric tree $\mathfrak{T}^\sigma_X$ is said to be open$_\ddagger$ just in case it contains a branch which is open$_\ddagger$, i.e., just in case $B_f(X_\sigma) \in O^\ddagger_M$ for some $B_f(X_\sigma) \in \mathfrak{T}^\sigma_X$. We write $O^\ddagger_M(X_\sigma)$ just in case $\mathfrak{T}^\sigma_X$ is open$_\ddagger$, and $C^\ddagger_M(X_\sigma)$ otherwise. In this sense, branch closure conditions induce closure conditions for AD sentences. These closure conditions can be used to define L_T valuations in the expected manner. That is:

$$\mathcal{V}^\ddagger_M(\sigma) = \begin{cases} \mathbf{a} & O^\ddagger_M(A_\sigma) \text{ and } C^\ddagger_M(D_\sigma); \\ \mathbf{b} & O^\ddagger_M(A_\sigma) \text{ and } O^\ddagger_M(D_\sigma); \\ \mathbf{n} & C^\ddagger_M(A_\sigma) \text{ and } C^\ddagger_M(D_\sigma); \\ \mathbf{d} & C^\ddagger_M(A_\sigma) \text{ and } O^\ddagger_M(D_\sigma). \end{cases} \tag{5.20}$$

Inducing valuations from branch closure conditions as in (5.20), we see that the following question, raised by Melvin Fitting, becomes highly relevant:

> Now the issue is: what closure conditions do we want to impose on a set S of signed statements to reflect our understanding of language and truth? (Fitting, [16, p80])

In this paper, we will only be concerned with the • closure conditions for branches. A branch B is contained in $C^\bullet_M$, just in case:

1) B contains X_σ with $X_\sigma \in At^*_M$ and $X_\sigma \notin w_M$, or

[13]The definition of Branch_M depends on M for the same reasons as EXP_M does.

2) B contains both A_σ and D_σ for some $\sigma \in Sen(L_T)$.

$\mathcal{V}^\bullet$ is a 4-valued theory of truth that will be used, below, to define $\mathcal{K}^+$ via MCG. $\mathcal{V}^\bullet$ is not an SK_4 theory, which is testified by the following observations:

$$\mathcal{V}^\bullet_M(T(\tau)) = \mathcal{V}^\bullet_M(\neg T(\tau)) = \mathbf{b}, \qquad \mathcal{V}^\bullet_M(T(\tau) \vee \neg T(\tau)) = \mathbf{a}$$

Despite the fact that it does not have a compositional semantics, I regard $\mathcal{V}^\bullet$ to be an interesting theory of truth; the $\bullet$ closure conditions represent an assertoric norm with a strong intuitive appeal: in asserting or denying a sentence you may never become committed to 1) an assertoric act which conflicts with the non-semantic facts, or 2) to contradict yourself. Or, in a catchy slogan: thou shalt respect the world and thou shalt not contradict thyself!

In fact, although hidden by its present definition, $\mathcal{V}^\bullet_M$ is a familiar valuation. As we will show next, $\mathcal{V}^\bullet_M$ is equivalent to Kripke's "modal fixed point valuation" $\mathcal{K}^{\mathbf{4}}_M$, which he obtained by quantifying over[14] $\mathbf{FP}_M$. $\mathcal{K}^{\mathbf{4}}_M$ is defined as follows, where the quantifiers range over $\mathbf{FP}_M$.

- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = \mathbf{a} \Leftrightarrow \exists V_M : V_M(\sigma) = \mathbf{a}$ and $\nexists V_M : V_M(\sigma) = \mathbf{d}$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = \mathbf{b} \Leftrightarrow \exists V_M : V_M(\sigma) = \mathbf{a}$ and $\exists V_M : V_M(\sigma) = \mathbf{d}$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = \mathbf{n} \Leftrightarrow \nexists V_M : V_M(\sigma) = \mathbf{a}$ and $\nexists V_M : V_M(\sigma) = \mathbf{d}$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = \mathbf{d} \Leftrightarrow \nexists V_M : V_M(\sigma) = \mathbf{a}$ and $\exists V_M : V_M(\sigma) = \mathbf{d}$

In order to prove that $\mathcal{V}^\bullet = \mathcal{K}^{\mathbf{4}}$ we need some definitions, which are all modifications of notions defined, amongst others, in [16].

Definition 5.12 Saturated sets, upwards closure
Let S be a set of AD sentences. We say that S is *downwards saturated* just in case:

$$\text{type of } X_\sigma \text{ is } \sqcup \Rightarrow (X_\sigma \in S \Rightarrow \Pi(X_\sigma) \cap S \neq \varnothing)$$
$$\text{type of } X_\sigma \text{ is } \sqcap \Rightarrow (X_\sigma \in S \Rightarrow \Pi(X_\sigma) \subseteq S)$$

This notion of an *upwards saturated* set is defined dually. That is, S is upwards saturated just in case:

$$\text{type of } X_\sigma \text{ is } \sqcup \Rightarrow (X_\sigma \in S \Leftarrow \Pi(X_\sigma) \cap S \neq \varnothing)$$
$$\text{type of } X_\sigma \text{ is } \sqcap \Rightarrow (X_\sigma \in S \Leftarrow \Pi(X_\sigma) \subseteq S)$$

Every set of AD sentences S has an *upwards closure* $S^{\Uparrow}$, i.e., a smallest set of AD sentences which extends S and which is upwards saturated[15]. □

[14] Remember that $\mathbf{FP}_M$ is the set of all 2 and 3 valued (Strong Kleene) fixed point valuations over M, as defined by Definition 5.3.

[15] The notions of downwards and upwards saturation are closely related to the notions of downwards and upwards saturation as defined by [16]. However, an important (and the only) difference between Fitting's notions and ours is that Fitting's notions are defined with respect to the assertoric rules for L only, i.e., in his definition Fitting does not treat the rules for truth not on par with the other rules. Likewise, the other notions defined in this section are inspired by [16] and differ from Fitting's notions only in the aspect just indicated. For a proof of the claim that every set of AD sentences has an upwards closure, see [16].

Definition 5.13 $\mathbf{FP}_M$ sets and associated valuations
Let S be a set of AD sentences. We say that S is an $\mathbf{FP}_M$ *set* just in case:

1. $\forall \sigma \in Sen(L_T)$: $A_\sigma \in S \Rightarrow D_\sigma \notin S$.
2. S is downwards and upwards saturated.
3. $w_M \subseteq S$.

An $\mathbf{FP}_M$ set S is a notational variant of the associated fixed point valuation, $V_M^S : Sen(L_T) \to \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$:

i) $A_\sigma \in S \Leftrightarrow V_M^S(\sigma) = \mathbf{a}$.

ii) $D_\sigma \in S \Leftrightarrow V_M^S(\sigma) = \mathbf{d}$.

iii) $\{A_\sigma, D_\sigma\} \cap S = \varnothing \Leftrightarrow V_M^S(\sigma) = \mathbf{n}$.

On the other hand, every fixed point valuation $V_M \in \mathbf{FP}_M$ corresponds[16], via i), ii) and iii) to an $\mathbf{FP}_M$ set S. □

Before we prove that $\mathcal{V}^\bullet = \mathcal{K}^\mathbf{4}$ it is instructive to comment on our proof strategy, which is a modification of the soundness and completeness proofs for signed tableaux systems. Consider the assertoric rules for $\wedge$ and $\neg$ for a propositional language $\mathcal{L}_P$ under the usual closure conditions: a branch[17] is closed just in case it contains, for some sentence σ of $\mathcal{L}_P$, both A_σ and D_σ and a tableau for X_σ is closed just in case all its branches are closed. This specifies a sound and complete signed tableau proof system with respect to the classical semantics of $\mathcal{L}_P$: a sentence σ of $\mathcal{L}_P$ is true in every $\mathcal{L}_P$ valuation just in case D_σ has a closed tableau. Soundness is proved by observing that if D_σ has a closed tableau, there is no $\mathcal{L}_P$ valuation in which σ is false. Completeness is proved by showing that if D_σ does not have a closed tableau, we can take an open branch and transform it into an $\mathcal{L}_P$ valuation which renders σ false.

We use our branches and assertoric trees to induce semantic valuations; our $\bullet$ closure conditions are defined relative to a ground model M. The role that is played by the classical valuations in the $\mathcal{L}_P$ case is, in our case, played by a (3-valued Strong Kleene) fixed point. If all the branches of $\mathfrak{T}_A^\sigma$ are closed$_\bullet$, there is no fixed point in which σ is valuated as $\mathbf{a}$. Similarly, if all the branches of $\mathfrak{T}_D^\sigma$ are closed$_\bullet$, there is no fixed point in which σ is valuated as $\mathbf{d}$. On the other hand, if $\mathfrak{T}_A^\sigma$ has an open$_\bullet$ branch, we can convert this branch into a fixed point which valuates σ as $\mathbf{a}$. Similarly for the case when $\mathfrak{T}_D^\sigma$ has an open$_\bullet$ branch. Let us now turn to the proof which makes these remarks precise.

Theorem 5.4 $\mathcal{V}_M^\bullet = \mathcal{K}_M^\mathbf{4}$
Let B be an open$_\bullet$ branch of A_σ. Then, $(B \cup w_M)^{\Uparrow}$, i.e., the upwards closure of $B \cup w_M$, is an $\mathbf{FP}_M$ set which contains A_σ. From this, it follows that:

$$O_M^\bullet(\mathfrak{T}_A^\sigma) \Rightarrow \exists V_M \in \mathbf{FP}_M : V_M(\sigma) = \mathbf{a}$$

[16] Where every $V_M \in \mathbf{FP}_M$ is thought of as having range $\{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$.

[17] The notion of a branch in this setting is slightly different from our definition of a branch. In fact, we use 'branch' to denote what is more commonly called 'completed branch'. Likewise, the notion of an assertoric tree differs from that of a tableau.

And, similarly, we get that:

$$O^{\bullet}_M(\mathfrak{T}^{\sigma}_D) \Rightarrow \exists V_M \in \mathbf{FP}_M : V_M(\sigma) = \mathbf{d}$$

On the other hand, let $\sigma \in Sen(L_T)$ and let $V_M \in \mathbf{FP}_M$ be such that $V_M(\sigma) = \mathbf{a}$. Let S be the $\mathbf{FP}_M$ set associated with V_M. Per definition, $A_\sigma \in S$. Let f be any strategy for player $\sqcup$ which is such that, for every $X_\sigma \in S$ of type $\sqcup$, $f(X_\sigma) \in S$. It follows that $B_f(A_\sigma)$ is open$_\bullet$. Similar remarks apply to $V_M(\sigma) = \mathbf{d}$ and D_σ. Hence, we get that

$$\exists V_M \in \mathbf{FP}_M : V_M(\sigma) = \mathbf{a} \Rightarrow O^{\bullet}_M(\mathfrak{T}^{\sigma}_A)$$

$$\exists V_M \in \mathbf{FP}_M : V_M(\sigma) = \mathbf{d} \Rightarrow O^{\bullet}_M(\mathfrak{T}^{\sigma}_D)$$

From the four established equations, it follows that $\mathcal{V}^{\bullet}_M = \mathcal{K}^{\mathbf{4}}_M$. □

Although $\mathcal{V}^{\bullet}_M$ is not defined via MCG, it is clearly defined using only notions that "belong" to the MCG. In the next subsection, we will use $\mathcal{V}^{\bullet}_M$ to define closure conditions which define $\mathcal{K}^{+}$. Doing so, we obtain a definition of $\mathcal{K}^{+}$ which is, in an important sense, not parasitic on Kripke's framework for truth.

5.5.2 Using $\mathcal{V}^{\bullet}$ to define $\mathcal{K}^{+}$

We will prove that $\mathcal{K}^{+}$ can be induced from closure conditions that are defined in terms of the notion of *strong* $\mathcal{V}^{\bullet}_M$ *correctness*. We say that X_σ is *strong* $\mathcal{V}^{\bullet}_M$ correct just in case:

$$X = A \& \mathcal{V}^{\bullet}_M(\sigma) = \mathbf{a} \quad \text{or} \quad X = D \& \mathcal{V}^{\bullet}_M(\sigma) = \mathbf{d}$$

We say that X_σ is *weak* $\mathcal{V}^{\bullet}_M$ *correct* just in case X_σ is $\mathcal{V}^{\bullet}_M$ correct in the sense of Definition 5.11. That is, X_σ is *weak* $\mathcal{V}^{\bullet}_M$ *correct* just in case:

$$X = A \& \mathcal{V}^{\bullet}_M(\sigma) \in \{\mathbf{a}, \mathbf{b}\} \quad \text{or} \quad X = D \& \mathcal{V}^{\bullet}_M(\sigma) \in \{\mathbf{d}, \mathbf{b}\}$$

Corresponding to the weak and strong notion of $\mathcal{V}^{\bullet}_M$ correctness, we define the *weak* and *strong closure conditions* as follows. With $\mathsf{exp} = \{y_n\}_{n \in \mathbb{N}}$, we let:

$$\mathsf{exp} \in O^{st}_M \Leftrightarrow \exists n \forall m > n : y_m \text{ is strong } \mathcal{V}^{\bullet}_M \text{ correct} \tag{5.21}$$

$$\mathsf{exp} \in O^{we}_M \Leftrightarrow \exists n \forall m > n : y_m \text{ is weak } \mathcal{V}^{\bullet}_M \text{ correct} \tag{5.22}$$

We will show that the valuation function induced by the strong closure conditions, i.e., $\mathcal{V}^{st}_M$, is equal to $\mathcal{K}^{+}_M$. Before we do so, however, we first sketch the rationale of the definition of $\mathcal{K}^{+}_M$ in terms of strong $\mathcal{V}^{\bullet}_M$ correctness.

For sure, if we have that $\mathcal{K}^{+}_M(\sigma) = \mathbf{a}$, we have that $\mathcal{V}^{\bullet}_M(\sigma) = \mathbf{a}$. For, if $\mathcal{K}^{+}_M(\sigma) = \mathbf{a}$, there is a (3-valued Strong Kleene) fixed point which valuates σ as $\mathbf{a}$ and also, there is no fixed point which valuates σ as $\mathbf{d}$. Similarly, $\mathcal{K}^{+}_M(\sigma) = \mathbf{d}$ implies that $\mathcal{V}^{\bullet}_M(\sigma) = \mathbf{d}$. The converses of these implications do not hold, however. For instance, we have that:

$$\mathcal{V}^{\bullet}_M(\neg T(\lambda) \vee T(\tau)) = \mathbf{a} \qquad \mathcal{K}^{+}_M(\neg T(\lambda) \vee T(\tau)) = \mathbf{n}$$

$$\mathcal{V}^{\bullet}_M(\neg T(\tau) \wedge T(\tau)) = \mathbf{d} \qquad \mathcal{K}^{+}_M(\neg T(\tau) \wedge T(\tau)) = \mathbf{n}$$

Although $A_{\neg T(\lambda)\vee T(\tau)}$ and $D_{\neg T(\tau)\wedge T(\tau)}$ are strong $\mathcal{V}^{\bullet}_M$ correct, none of their immediate AD subsentences is strong $\mathcal{V}^{\bullet}_M$ correct. This ensures, as is readily noticed, that $\mathcal{V}^{st}_M(\neg T(\lambda)\vee T(\tau)) = \mathcal{V}^{st}_M(\neg T(\tau)\wedge T(\tau)) = \mathbf{n}$, mimicking the judgment of $\mathcal{K}^{+}_M$ with respect to these sentences. More generally, the definition of $\mathcal{V}^{st}_M$ ensures that, for AD sentences which are "unstable" strong $\mathcal{V}^{\bullet}_M$ correct—i.e., ultimately, they depend on a combination of AD sentences which are not strong $\mathcal{V}^{\bullet}_M$ correct—player $\sqcup$ does not have a strategy which ensures that his expansion ends up in O^{st}_M. In order to prove that $\mathcal{V}^{st}_M = \mathcal{K}^{+}_M$, we will evoke the following three lemma's.

Lemma 5.1 • openess is preserved downwards
By the phrase '• openess is preserved downwards', we mean that:

$$\text{type of } X_\sigma \text{ is } \sqcup \Rightarrow (O^{\bullet}_M(X_\sigma) \Rightarrow \exists Y_\alpha \in \Pi(X_\sigma) : O^{\bullet}_M(Y_\alpha))$$
$$\text{type of } X_\sigma \text{ is } \sqcap \Rightarrow (O^{\bullet}_M(X_\sigma) \Rightarrow \forall Y_\alpha \in \Pi(X_\sigma) : O^{\bullet}_M(Y_\alpha))$$

Proof: This follows immediate from an inspection of the • closure conditions and the observation that the branches which constitute the tree of an immediate AD subsentence of X_σ are subsets of the branches which constitute the tree of X_σ. □

Lemma 5.2 $\mathcal{V}^{st}_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$ is an SK_3 theory
It is clear that the strong closure conditions satisfy SJC and WRC and so, by the (corollary to the) first stable judgement Theorem, they define a Strong Kleene theory. The point of this lemma then, is to show that $\mathcal{V}^{st}_M$ is 3 valued. To do so, we proceed as in Section 5.4. Suppose that $O^{st}_M(A_\sigma)$ and let f be a strategy of $\mathcal{F}$ which ensures that the expansion of A_σ ends up in O^{st}_M. The *mirror strategy* of f, g_f (see Section 5.4), testifies that $C^{st}_M(D_\sigma)$. For, if an AD sentence X_α is strong $\mathcal{V}^{\bullet}_M$ correct, then its inverse X^{-1}_α is not. □

Our proof of the fact that $\mathcal{V}^{st}_M = \mathcal{K}^{+}_M$ will exploit a further lemma, which invokes the notion of a *totally strong $\mathcal{V}^{\bullet}_M$ correct expansion*. An expansion is said to be totally strong $\mathcal{V}^{\bullet}_M$ correct just in case *all* its terms are strong $\mathcal{V}^{\bullet}_M$ correct. Here is the lemma:

Lemma 5.3 $\forall g \in \mathcal{G} : \mathsf{exp}(X_\sigma, f', g) \in O^{st}_M \Leftrightarrow \forall g \in \mathcal{G} : \mathsf{exp}(X_\sigma, f', g)$ is totally strong $\mathcal{V}^{\bullet}_M$ correct
Proof: The right to left direction is trivial. For the converse direction, let f' be a strategy which testifies that $O^{st}_M(X_\sigma)$, i.e.,:

$$\forall g \in \mathcal{G} : \mathsf{exp}(X_\sigma, f', g) \in O^{st}_M$$

Let $g' \in \mathcal{G}$. We have to show that $\mathsf{exp}' = \mathsf{exp}(X_\sigma, f', g')$ is totally strong $\mathcal{V}^{\bullet}_M$ correct. As $\mathsf{exp}' \in O^{st}_M$, exp' contains a first strong $\mathcal{V}^{\bullet}_M$ correct term (after which all other terms are strong $\mathcal{V}^{\bullet}_M$ correct). We will prove by contraposition that this first term is equal to X_σ. Thus, assume that exp' contains a first strong $\mathcal{V}^{\bullet}_M$ correct term and that this term has a predecessor on exp' which is not strong $\mathcal{V}^{\bullet}_M$ correct. We assume, without loss of generality, that the first strong $\mathcal{V}^{\bullet}_M$ correct term has form A_α, the case where its form is D_α being similar. The predecessor of A_α on exp has one of the following six forms:

$$D_{\neg\alpha}, A_{\alpha\vee\beta}, A_{\alpha\wedge\beta}, A_{\forall x\phi(x)}, A_{\exists x\phi(x)}, A_{T(\overline{\alpha})}$$

We only prove the claim for the cases where the predecessor of A_α is $A_{\alpha\vee\beta}$ or $A_{\alpha\wedge\beta}$, as the other four cases are either trivial or similar to the two cases that we will discuss.

Predecessor of A_α is $A_{\alpha\vee\beta}$. As A_α is strong $\mathcal{V}^\bullet_M$ correct, we have that $\mathcal{V}^\bullet_M(\alpha) = \mathbf{a}$. Hence, there is a (3-valued Strong Kleene) fixed point in which α is valuated as $\mathbf{a}$ and no fixed point in which α is valuated as $\mathbf{d}$. In the fixed point in which α is valuated as $\mathbf{a}$, $\alpha\vee\beta$ is also valuated as $\mathbf{a}$. Thus, $\mathcal{V}^\bullet_M(\alpha\vee\beta) \in \{\mathbf{a},\mathbf{b}\}$. Suppose that $\mathcal{V}^\bullet_M(\alpha\vee\beta) = \mathbf{a}$. This gives a contradiction with the assumption that A_α is the first strong $\mathcal{V}^\bullet_M$ correct element on exp'. Thus, suppose that $\mathcal{V}^\bullet_M(\alpha\vee\beta) = \mathbf{b}$. Per definition of $\mathcal{V}^\bullet_M$, we get $O^\bullet_M(D_{\alpha\vee\beta})$. From Lemma 5.1, we get that $O^\bullet_M(D_\alpha)$ and $O^\bullet_M(D_\beta)$. From $O^\bullet_M(D_\alpha)$ it follows, by Theorem 5.4, that there is a fixed point in which α is valuated as $\mathbf{d}$. This gives a contradiction with strong $\mathcal{V}^\bullet_M$ correctness of A_α.

Predecessor of A_α is $A_{\alpha\wedge\beta}$. As A_α is strong $\mathcal{V}^\bullet_M$ correct, we have that $\mathcal{V}^\bullet_M(\alpha) = \mathbf{a}$. Further, strategy f' (by considering the mirror strategy of f' as in the proof of Lemma 5.2) testifies that $\mathcal{V}^{st}_M(\alpha\wedge\beta) = \mathcal{V}^{st}_M(\alpha) = \mathcal{V}^{st}_M(\beta) = \mathbf{a}$. From the fact that $\mathcal{V}^{st}_M(\alpha\wedge\beta) = \mathbf{a}$, it follows that there is a 3 valued fixed point (namely, $\mathcal{V}^{st}_M$) in which $\alpha\wedge\beta$ is valuated as $\mathbf{a}$. Hence, from Theorem 5.4, it follows that $\mathcal{V}^\bullet_M(\alpha\wedge\beta) \in \{\mathbf{a},\mathbf{b}\}$. Suppose that $\mathcal{V}^\bullet_M(\alpha\wedge\beta) = \mathbf{a}$. This gives a contradiction with the assumption that A_α is the first strong $\mathcal{V}^\bullet_M$ correct element on exp'. Thus, suppose that $\mathcal{V}^\bullet_M(\alpha\wedge\beta) = \mathbf{b}$. From Lemma 5.1, we get that $O^\bullet_M(A_\beta)$. Further, from $\mathcal{V}^\bullet_M(\alpha) = \mathbf{a}$ it follows, per definition, that $C^\bullet_M(D_\alpha)$. Similarly, from $\mathcal{V}^\bullet_M(\alpha\wedge\beta) = \mathbf{b}$ we get, per definition, that $O^\bullet_M(D_{\alpha\wedge\beta})$. From $O^\bullet_M(D_{\alpha\wedge\beta})$ and $C^\bullet_M(D_\alpha)$ it follows that $O^\bullet_M(D_\beta)$ and so $\mathcal{V}^\bullet_M(\beta) = \mathbf{b}$. Hence A_β is not strong $\mathcal{V}^\bullet_M$ correct. Now, let $g'' \in \mathcal{G}$ be the strategy that is defined just like g' except for the fact that $g'(A_{\alpha\wedge\beta}) = A_\alpha$, whereas $g''(A_{\alpha\wedge\beta}) = A_\beta$. Let $\mathsf{exp}'' = \mathsf{exp}(X_\sigma, f', g'')$ be the expansion of X_σ induced by f' and g'' and note that $A_{\alpha\wedge\beta}$ occurs on exp''. Let Y_γ be the first element of type $\sqcup$ which occurs on exp'' after $A_{\alpha\wedge\beta}$ such that $|\Pi(Y_\gamma)| > 1$. If there is no such element if follows, from Lemma 5.1, that for every element Z_θ which occurs on exp'', we have that $\mathcal{V}^\bullet_M(Z_\theta) = \mathbf{b}$. Observe that this contradicts with the assumption that strategy f' guarantees that for *every* g, $\mathsf{exp}(X_\sigma, f', g)$ ends up in O^{st}_M. Thus, let Y_γ be as indicated. From Lemma 5.1, it follows that $\Pi(Y_\gamma)$ contains at least one element, say Y_δ, such that $O^\bullet_M(Y_\delta)$. Moreover, from the definition of f', it follows that f' has to pick an $Y_\delta \in \Pi(Y_\gamma)$ such that $O^\bullet_M(Y_\delta)$. For suppose not, i.e., suppose that $f'(Y_\gamma) = Y_{\delta'}$ such that $C^\bullet_M(Y_{\delta'})$. According to Theorem 5.4, this means that there is no 3 valued fixed point which contains $Y_{\delta'}$. On the other hand, from the definition of f' and the assumption that $f'(Y_\gamma) = Y_{\delta'}$, it follows that there is a 3 valued fixed point (namely, $\mathcal{V}^{st}_M$) which contains $Y_{\delta'}$. Thus, $f'(Y_\gamma) = Y_\delta$ for some Y_δ such that $O^\bullet_M(Y_\delta)$. From Lemma 5.1, the fact that $\mathcal{V}^\bullet_M(\alpha\wedge\beta) = \mathbf{b}$ and the fact that Y_γ is the first element on exp'' after $A_{\alpha\wedge\beta}$ for which player $\sqcup$ has to make a genuine choice, it follows that $O^\bullet_M(Y_\delta^{-1})$. Hence, we have that $\mathcal{V}^\bullet_M(\delta) = \mathbf{b}$. And so Y_δ is weak but not strong $\mathcal{V}^\bullet_M$ correct. We are now back were we started, with δ playing the role of β. We can repeat the argument, by looking at the first element which occurs on exp'' after Y_γ for which player $\sqcup$ has to make a genuine choice. By a similar argument, f' must allot a weak $\mathcal{V}^\bullet_M$ correct element to it. Hence, f' does not guarantee that for every g, $\mathsf{exp}(X_\sigma, f', g)$ ends up in O^{st}_M. □

Before we (finally) show that $\mathcal{V}^{st}_M = \mathcal{K}^+_M$, we first recall the definition of $\mathcal{K}^+_M$

in terms of the $\mathcal{K}^+$ closure conditions that are associated with the second stable judgement theorem. With $\mathsf{exp} = \{y_n\}_{n\in\mathbb{N}}$, these closure conditions are defined as follows:

$$\mathsf{exp} \in O_M^{\mathcal{K}^+} \Leftrightarrow \exists n \forall m > n : y_m \text{ is } \mathcal{K}_M^+ \text{ correct}$$

Theorem 5.5 $\mathcal{V}_M^{st} = \mathcal{K}_M^+$
Proof : It suffices to show that, for every AD sentence X_σ, it holds that:

$$O_M^{\mathcal{K}^+}(X_\sigma) \Leftrightarrow O_M^{st}(X_\sigma)$$

The left to right direction is immediate from the definition of $O_M^{\mathcal{K}^+}$ and O_M^{st}. Thus, assume that $O_M^{st}(X_\sigma)$. This means that there exists an $f \in \mathcal{F}$ such that for every $g \in \mathcal{G}$, $\mathsf{exp}(X_\sigma, f, g) \in O_M^{st}$. By Lemma 5.3, this means that every term that occurs on an expansion of X_σ that is induced by f, is strong $\mathcal{V}_M^\bullet$ correct. Hence, all elements of $B_f(X_\sigma)$, the branch of X_σ as induced by f, are strong $\mathcal{V}_M^\bullet$ correct. From this, it follows that the (3-valued Strong Kleene) fixed point valuation induced by $B_f(X_\sigma)^{\Uparrow}$, i.e., by the upwards closure of $B_f(X_\sigma)$, is compatible (see Section 5.3) with every fixed point valuation over M and hence is an intrinsic fixed point valuation, i.e., a member of $\mathbf{I}_M$ (see Definition 5.5). With S the $\mathbf{FP}_M$ set corresponding to $\mathcal{K}_M^+$, we get that $B_f(X_\sigma)^{\Uparrow} \subseteq S$, as $\mathcal{K}_M^+$ is *maximal* intrinsic. From $B_f(X_\sigma) \subseteq S$, it follows that $O_M^{\mathcal{K}^+}(X_\sigma)$. □

5.6 Generalized Strong Kleene theories of truth

5.6.1 Some GSK theories

As testified by the assertoric transfer theorem, GSK theories are typically build up from SK_3 and SK_4 theories which respect one another. In this subsection, we study the SK_3 and SK_4 theories that we defined thus far with an eye on the "respecting relation", i.e., the relation $<$ that was defined in Section 5.3. We then use the obtained results in combination with the assertoric transfer theorem to define GSK theories.

Proposition 5.3 Relating some SK_3 and SK_4 theories
We have that:

1. $\mathcal{K} < \mathcal{K}^+$
2. $\mathcal{K} < \mathcal{V}^{\blacklozenge}$
3. $\mathcal{K}^+ < \mathcal{V}^{we}$
4. $\mathcal{K}^+ \not< \mathcal{V}^{\blacklozenge}$

Proof:
1. Folklore.
2. Suppose that $\mathcal{K}_M(\sigma) = \mathbf{a}$. Thus, by Proposition 5.1, $O_M^{gr}(A_\sigma)$, from which it follows that $O_M^{\blacklozenge}(A_\sigma)$. Let $f' \in \mathcal{F}$ be a strategy which testifies that $O_M^{gr}(A_\sigma)$, i.e. f' is such that $\mathsf{exp}(A_\sigma, f', g) \in G_M^{cor}$ for every $g \in \mathcal{G}$. By arguments familiar from Section 5.4, it follows that the mirror strategy of f', $g_{f'}$, ensures that $\mathsf{exp}(D_\sigma, f, g_{f'}) \in G_M^{inc}$ for every $f \in \mathcal{F}$. From this, it follows that $C_M^{\blacklozenge}(D_\sigma)$ and so we have that $\mathcal{V}_M^{\blacklozenge}(\sigma) = \mathbf{a}$, which is what we had to show.
3. Similar to the proof of 2., now exploiting the representation of $\mathcal{K}^+$ as $\mathcal{V}^{st}$ (rather than the representation of $\mathcal{K}$ as $\mathcal{V}^{gr}$) and the relation between the strong and weak closure conditions.

4. It is well-known (and easily established) that a Tautologyteller, i.e. a sentence $T(\eta) \vee \neg T(\eta)$ where $I(\eta) = T(\eta) \vee \neg T(\eta)$ is valuated as $\mathbf{a}$ by $\mathcal{K}^+_M$. It is left to the reader to observe that $\mathcal{V}^{\blacklozenge}_M(T(\eta) \vee \neg T(\eta)) = \mathbf{b}$. □

The relation between $\mathcal{K}$ and $\mathcal{K}^+$ allow us to combine these theories, in accordance with the assertoric transfer theorem, into a five valued GSK theory, $\mathbb{V}^{\mathbf{5+}}$. Similarly, the relation between $\mathcal{K}$ and $\mathcal{V}^{\blacklozenge}$ allows us to combine these theories into a six valued GSK theory, $\mathbb{V}^{\mathbf{6}}$ in accordance with the assertoric transfer theorem. That is:

$$\mathcal{V}^{\mathbf{5+}}_M(\sigma) = \begin{cases} \mathbf{a}_g, & \mathcal{K}(\sigma) = \mathbf{a}; \\ \mathbf{a}_i, & \mathcal{K}^+(\sigma) = \mathbf{a}, \mathcal{K}(\sigma) = \mathbf{n}; \\ \mathbf{e}, & \mathcal{K}^+(\sigma) = \mathbf{n}; \\ \mathbf{d}_i, & \mathcal{K}^+(\sigma) = \mathbf{d}, \mathcal{K}(\sigma) = \mathbf{n}; \\ \mathbf{d}_g, & \mathcal{K}(\sigma) = \mathbf{d}. \end{cases}$$

$$\mathcal{V}^{\mathbf{6}}_M(\sigma) = \begin{cases} \mathbf{a}_g, & \mathcal{K}(\sigma) = \mathbf{a}; \\ \mathbf{x}_u, & \mathcal{K}(\sigma) = \mathbf{n}, \mathcal{V}^{\blacklozenge}_M(\sigma) = \mathbf{x}; \\ \mathbf{d}_g, & \mathcal{K}(\sigma) = \mathbf{d}. \end{cases}$$

Here, the subscripts g, u, i that are attached to the assertoric values stand for, respectively, *grounded*, *ungrounded* and *intrinsic*. The value $\mathbf{e}$ that is recognized by $\mathcal{V}^{\mathbf{5+}}$ indicates that we call sentences that are valuated as such *extrinsic*.

Also, the relation between $\mathcal{K}^+$ and $\mathcal{V}^{we}$ allows us to combine these theories into a six valued GSK theory, say $\mathbb{V}^{\mathbf{6+}}$. As $\mathcal{K}$ respects $\mathbb{V}^{\mathbf{6+}}$, a further application of the assertoric transfer theorem allows us to combine $\mathcal{K}$ and $\mathbb{V}^{\mathbf{6+}}$ into the eight valued GSK theory, $\mathbb{V}^{\mathbf{8+}}$, that was discussed in the introduction. Equivalently, $\mathbb{V}^{\mathbf{8+}}$ can be defined in terms of $\mathcal{V}^{\mathbf{5+}}$ and $\mathcal{V}^{we}$:

$$\mathbb{V}^{\mathbf{8+}}_M(\sigma) = \begin{cases} \mathcal{V}^{\mathbf{5+}}_M(\sigma), & \mathcal{V}^{\mathbf{5+}}_M(\sigma) \neq \mathbf{e}; \\ \mathbf{x}_e & \mathcal{V}^{\mathbf{5+}}_M(\sigma) = \mathbf{e}, \mathcal{V}^{we}_M(\sigma) = \mathbf{x}. \end{cases} \tag{5.23}$$

The subscript e, attached to assertoric values, indicates that a sentence is extrinsic. The definition of $\mathbb{V}^{\mathbf{8+}}$ as given by (5.23) clearly indicates that $\mathbb{V}^{\mathbf{8+}}$ gives a compositional account of the sentences that are called 'extrinsic' (i.e. valuated as $\mathbf{e}$) by $\mathcal{V}^{\mathbf{5+}}$. Below we display the Hasse diagrams associated with $\mathbb{V}^{\mathbf{5+}}$, $\mathbb{V}^{\mathbf{6}}$ and $\mathbb{V}^{\mathbf{8+}}$.

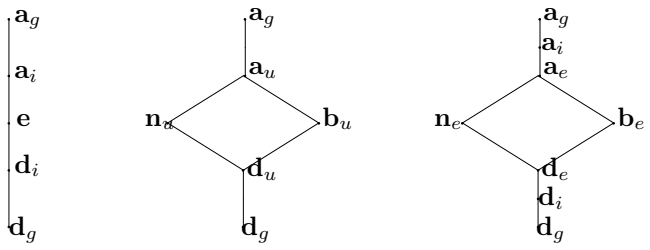


Figure 5.4: $\mathbb{V}^{\mathbf{5+}}$ $(\mathcal{K}, \mathcal{K}^+)$, $\mathbb{V}^{\mathbf{6}}$ $(\mathcal{K}, \mathcal{V}^{\blacklozenge})$ and $\mathbb{V}^{\mathbf{8+}}$ $(\mathcal{K}, \mathcal{K}^+, \mathcal{V}^{we})$.

The GSK compositionality of these theories was explained in Section 2: the only difference with a SK theory is that negation swaps $\mathbf{a}_x$ with $\mathbf{d}_x$, where

$x \in \{g, u, i, e\}$.

Let us illustrate the necessity of the fulfillment of the conditions of the assertoric transfer theorem for its definition of a GSK theory. Suppose that we define a six valued theory, call it $\mathbb{V}^{six}$, in terms of the theories $\mathcal{K}^+$ and $\mathcal{V}^{\blacklozenge}$ ($\mathcal{V}^{\blacklozenge}$ does not respect $\mathcal{K}^+$, see Proposition 5.3) according to the definition that is employed by the assertoric transfer theorem:

$$\mathbb{V}^{six}_M(\sigma) = \begin{cases} \mathbf{a}_1, & \mathcal{K}^+(\sigma) = \mathbf{a}; \\ \mathbf{x}_0, & \mathcal{K}^+(\sigma) = \mathbf{n}, \mathcal{V}^{\blacklozenge}_M(\sigma) = \mathbf{x}; \\ \mathbf{d}_1, & \mathcal{K}^+(\sigma) = \mathbf{d}. \end{cases}$$

To see that $\mathbb{V}^{six}$ is not a GSK theory, let $I(\eta) = T(\eta) \vee \neg T(\eta)$ be a Tautologyteller and observe that:

$$\mathcal{K}^+_M(T(\eta)) = \mathbf{a}, \quad \mathcal{K}^+_M(\neg T(\lambda)) = \mathbf{n} \quad \mathcal{V}^{\blacklozenge}_M(T(\eta)) = \mathbf{b}, \quad \mathcal{V}^{\blacklozenge}_M(\neg T(\lambda)) = \mathbf{n}$$

From the compositionality of $\mathcal{K}^+$ and $\mathcal{V}^{\blacklozenge}$, it follows that $\mathcal{K}^+_M(T(\eta) \wedge \neg T(\lambda)) = \mathbf{n}$ and that $\mathcal{V}^{\blacklozenge}_M(T(\eta) \wedge \neg T(\lambda)) = \mathbf{d}$. Hence, according to the definition of $\mathbb{V}^{six}$, we get that:

$$\mathbb{V}^{six}_M(T(\eta)) = \mathbf{a}_1, \quad \mathbb{V}^{six}_M(\neg T(\lambda)) = \mathbf{n}_0, \quad \mathbb{V}^{six}_M(T(\eta) \wedge \neg T(\lambda)) = \mathbf{d}_0$$

This shows that $\mathbb{V}^{six}$ is not a GSK theory. For it were a GSK theory, its compositionality would be described by a six valued GSK lattice, according to which $\mathbb{V}^{six}_M(T(\eta)) = \mathbf{a}_1$ together with $\mathbb{V}^{six}_M(\neg T(\lambda)) = \mathbf{n}_0$ implies that $\mathbb{V}^{six}_M(T(\eta) \wedge \neg T(\lambda)) = \mathbf{n}_0$.

Thus far, we defined three GSK theories, of cardinality 5, 6 and 8 respectively. What about GSK theories of other cardinalities? In order to answer that question, we define some further SK_3 and SK_4 theories which are suitably related to the SK theories that are already defined. In particular, we will be interested in theories that respect $\mathcal{K}^+$. In order to define such theories, the next section takes a closer look at the closure conditions of $\mathcal{K}^+$.

5.6.2 A closer look at the closure conditions of $\mathcal{K}^+$

Recall the definition of $\mathcal{K}^+$ in terms of the $\mathcal{K}^+$ closure conditions that are associated with the second stable judgement theorem. Here they are:

$$\mathsf{exp} \in O^{\mathcal{K}^+}_M \Leftrightarrow \exists n \forall m > n : y_m \text{ is } \mathcal{K}^+_M \text{ correct}$$

Also, recall the definition of $\mathcal{K}$ in terms of the groundedness closure conditions:

$$O^{gr}_M = G^{cor}_M$$

From the fact that $\mathcal{K}^+$ respects $\mathcal{K}$, it follows that $O^{gr}_M = G^{cor}_M \subseteq O^{\mathcal{K}^+}_M$. In fact, the subset relation is strict, as is testified by the following expansion of a Tautologyteller:

$$A_{T(\eta)\vee\neg T(\eta)}, A_{T(\eta)}, A_{T(\eta)\vee\neg T(\eta)}, A_{T(\eta)}, A_{T(\eta)\vee\neg T(\eta)}, \ldots$$

The depicted expansion of $A_{T(\eta)\vee\neg T(\eta)}$ testifies that $O^{\mathcal{K}^+}_M$ contains, in terms of Section 5.4, expansions which are ungrounded and stable$_A$. Similarly, the following expansion of the Contradictionteller testifies that $O^{\mathcal{K}^+}_M$ contains expansions

which are ungrounded and stable$_D$:

$$D_{T(\theta)\wedge\neg T(\theta)}, D_{T(\theta)}, D_{T(\theta)\wedge\neg T(\theta)}, D_{T(\theta)}, D_{T(\theta)\wedge\neg T(\theta)}, \dots$$

Two important characteristics of expansions in O^{gr}_M is that they are all stable (i.e., either stable$_A$ or stable$_D$), and also, that their set of terms is finite. The expansions in $O^{\mathcal{K}^+}_M$ do not (necessarily) share these characteristics, as the following example testifies. Let $C = \{c_n\}_{n\in\mathbb{N}}$ be a set of non-quotational constants which is interpreted as follows:

$$I(c_n) = \begin{cases} T(c_1) \vee \neg T(\lambda), & n = 0; \\ \neg(T(c_{n+1}) \wedge \neg T(\lambda)), & n \neq 0, n \text{ is odd}; \\ \neg(T(c_{n+1}) \vee \neg T(\lambda)), & n \neq 0, n \text{ is even}. \end{cases}$$

The following expansion of $A_{T(c_0)}$ is unstable, contains infinitely many distinct terms, and is an element of $O^{\mathcal{K}^+}_M$:

$$A_{T(c_0)}, A_{T(c_1)\vee\neg T(\lambda)}, A_{T(c_1)}, A_{\neg(T(c_2)\wedge\neg T(\lambda))}, D_{T(c_2)\wedge\neg T(\lambda)}, D_{T(c_2)}, \dots$$

$$\dots D_{\neg(T(c_3)\vee\neg T(\lambda))}, A_{T(c_3)\vee\neg T(\lambda)}, A_{T(c_3)}, A_{\neg(T(c_4)\wedge\neg T(\lambda))}, \dots \tag{5.24}$$

The fact that the expansion of $A_{T(c_0)}$ is unstable and contains infinitely many distinct terms is clear. Let us show that it is contained in $O^{\mathcal{K}^+}_M$. To do so, we show that each AD sentence on the expansion is $\mathcal{K}^+_M$ correct. Consider the AD sentence $A_{T(c_1)\vee\neg T(\lambda)}$. From the depicted expansion of $A_{T(c_0)}$ it follows that it is possible to valuate $T(c_1) \vee \neg T(\lambda)$ as $\mathbf{a}$ (for each sentence σ that is mentioned on (5.24), valuate σ as $\mathbf{a}$ when A_σ occurs on (5.24) and valuate σ as $\mathbf{d}$ when D_σ occurs on (5.24)). To show that $A_{T(c_1)\vee\neg T(\lambda)}$ is $\mathcal{K}^+_M$ correct, we have to show that that "letting" $T(c_1) \vee \neg T(\lambda)$ valuate as $\mathbf{a}$ by valuating $T(c_1)$ as $\mathbf{a}$ can be done in a non arbitrary manner, i.e., that there is no $V_M \in \mathbf{FP}_M$ which valuates $T(c_1)$ as $\mathbf{d}$. Suppose that $V_M \in \mathbf{FP}_M$ valuates $T(c_1)$ as $\mathbf{d}$. This means, as V_M respects the identity of truth, that V_M valuates $\neg(T(c_2) \wedge \neg T(\lambda))$ as $\mathbf{d}$, which means that V_M valuates $T(c_2) \wedge \neg T(\lambda)$ as $\mathbf{a}$. Hence, V_M has to valuate the Liar as $\mathbf{a}$, which gives a contradiction with V_M being an element of $\mathbf{FP}_M$. Hence, $A_{T(c_1)\vee\neg T(\lambda)}$ is $\mathcal{K}^+_M$ correct. In fact, an inspection of the expansion of $A_{T(c_0)}$ reveals that all elements of the expansion are $\mathcal{K}^+_M$ correct. Accordingly, the depicted expansion of $A_{T(c_0)}$ is contained in $O^{\mathcal{K}^+}_M$.

We thus showed that $O^{\mathcal{K}^+}_M$ contains ungrounded expansions which may be either stably$_A$, stably$_D$ or unstable. We use this observation to partition $O^{\mathcal{K}^+}_M$ into four subsets, using notions of Definition 5.10:

$$U^+_M = O^{\mathcal{K}^+}_M - G^{cor}_M$$

$$US^{A+}_M = U^+_M \cap US^A_M, \;\; US^{D+}_M = U^+_M \cap US^D_M, \;\; UU^+_M = U^+_M \cap UU_M$$

From the definitions just given, it readily follows that G^{cor}_M, US^{A+}_M, US^{D+}_M and UU^+_M are pairwise disjoint and that:

$$O^{\mathcal{K}^+}_M = G^{cor}_M \cup US^{A+}_M \cup US^{D+}_M \cup UU^+_M$$

We will first use this representation of $O^{\mathcal{K}^+}_M$ to define two SK_3 theories, which are called $\mathcal{V}^{\mathbf{3A}+}$ and $\mathcal{V}^{\mathbf{3D}+}$, and which respect $\mathcal{K}^+$. Theory $\mathcal{V}^{\mathbf{3A}+}$ is obtained by

modifying the $\mathcal{K}^+$ closure condition by declaring "as much stable$_A$ expansions open as possible", while $\mathcal{V}^{\mathbf{3D+}}$ is obtained by modifying the $\mathcal{K}^+$ closure condition by declaring "as much stable$_D$ expansions open as possible". To explain these phrases, we focus only on $\mathcal{V}^{\mathbf{3A+}}$, as the $\mathcal{V}^{\mathbf{3D+}}$ case is completely dual. Of course, not all stable$_A$ expansions can—with the purpose of defining a SK_3 theory which respects $\mathcal{K}^+$ in mind—be declared open: consider a stable$_A$ expansion that is contained in G^{inc}_M. On the other hand, it is unproblematic to declare the stable$_A$ expansions which end with an assertion of a Truthteller to be open:

$$A_{T(\tau)}, A_{T(\tau)}, A_{T(\tau)}, \ldots,$$

More generally, it is unproblematic to declare a stable$_A$ expansion to be open just in case its inverse (which is a stable$_D$ expansion per definition) is not contained in $O^{\mathcal{K}^+}_M$. This leads to the following definition of the closure conditions for $\mathcal{V}^{\mathbf{3A+}}$ and $\mathcal{V}^{\mathbf{3D+}}$:

$$O^{\mathbf{3A+}}_M = G^{cor}_M \cup (US^A_M - (US^{D+}_M)^{-1}) \cup US^{D+}_M \cup UU^+_M$$

$$O^{\mathbf{3D+}}_M = G^{cor}_M \cup (US^D_M - (US^{A+}_M)^{-1}) \cup US^{A+}_M \cup UU^+_M$$

Note that the set $(US^A_M - (US^{D+}_M)^{-1})$ consists of all ungrounded stable$_A$ expansions except for those whose inverse is contained in $O^{\mathcal{K}^+}_M$. As an example, $(US^A_M - (US^{D+}_M)^{-1})$ does not contain the following expansion of an assertion of the Contradictionteller:

$$A_{T(\theta)\wedge\neg T(\theta)}, A_{T(\theta)}, A_{T(\theta)\wedge\neg T(\theta)}, A_{T(\theta)}, \ldots \tag{5.25}$$

The following proposition states that $\mathcal{V}^{\mathbf{3A+}}$ and $\mathcal{V}^{\mathbf{3D+}}$ have the desired properties.

Proposition 5.4 **$\mathcal{V}^{\mathbf{3A+}}$ and $\mathcal{V}^{\mathbf{3D+}}$ are SK_3, $\mathcal{K}^+ < \mathcal{V}^{\mathbf{3A+}}$ and $\mathcal{K}^+ < \mathcal{V}^{\mathbf{3D+}}$.**
Proof: We illustrate only for $\mathcal{V}^{\mathbf{3A+}}$, the $\mathcal{V}^{\mathbf{3D+}}$ case being similar. Compare the closure conditions of $\mathcal{K}^+$ with those of $\mathcal{V}^{\mathbf{3A+}}$:

$$O^{\mathcal{K}^+}_M = G^{cor}_M \cup US^{A+}_M \cup US^{D+}_M \cup UU^+_M$$

$$O^{\mathbf{3A+}}_M = G^{cor}_M \cup (US^A_M - (US^{D+}_M)^{-1}) \cup US^{D+}_M \cup UU^+_M$$

Per definition, $(US^{D+}_M)^{-1}$ is such that $(US^{D+}_M)^{-1} \subseteq US^A_M$. Further, we have that $(US^{D+}_M)^{-1} \cap US^{A+}_M = \varnothing$. To see this, suppose that $\mathsf{exp} \in US^{A+}_M \subseteq O^{\mathcal{K}^+}_M$. From the $\mathcal{K}^+$ closure conditions, it follows that $\mathsf{exp}^{-1} \in US^D_M$ is contained in $C^{\mathcal{K}^+}_M$. Hence $\mathsf{exp}^{-1} \notin US^{D+}_M$ and so $(\mathsf{exp}^{-1})^{-1} = \mathsf{exp} \notin (US^{D+}_M)^{-1}$. Thus, we have that:

$$US^{A+}_M \subseteq (US^A_M - (US^{D+}_M)^{-1})$$

And so we have that $O^{\mathcal{K}^+}_M \subseteq O^{\mathbf{3A+}}_M$. From this, it follows that for every AD sentence X_σ, we have that:

$$O^{\mathcal{K}^+}_M(X_\sigma) \Rightarrow O^{\mathbf{3A+}}_M(X_\sigma) \tag{5.26}$$

Further, by an inspection of the closure conditions for $\mathcal{V}^{\mathbf{3A+}}$ we see that they respect SJC and WRC (from which it follows that $\mathcal{V}^{\mathbf{3A+}}$ is SK_3 or SK_4) and also, that:

$$(O^{\mathbf{3A+}}_M)^{-1} \subseteq C^{\mathbf{3A+}}_M \tag{5.27}$$

From (5.27), it now follows (by a mirror strategy argument familiar from Section 5.4) that $\mathcal{V}^{\mathbf{3A}+}$ is SK_3. Further, from (5.26) and (5.27) it follows that $\mathcal{V}^{\mathbf{3A}+}$ respects $\mathcal{K}^+$: suppose that $\mathcal{K}^+_M(\sigma) = \mathbf{a}$. Then we have that $O^{\mathcal{K}^+}_M(A_\sigma)$ and so, by (5.26), we get that $O^{\mathbf{3A}+}_M(A_\sigma)$. From (5.27) it follows that $C^{\mathbf{3A}+}_M(D_\sigma)$ and so we have that $\mathcal{V}^{\mathbf{3A}+}_M(\sigma) = \mathbf{a}$. □

As $\mathcal{V}^{\mathbf{3A}+}$ respects $\mathcal{K}^+$, the assertoric transfer theorem tells us that $\mathcal{V}^{\mathbf{3A}+}$ and $\mathcal{K}^+$ can be combined into a five valued GSK theory, call it $\mathbb{V}^{\mathbf{5}+}$. Further, as $\mathbb{V}^{\mathbf{5A}+}$ respects $\mathcal{K}$, another application of the assertoric transfer theorem to those theories delivers a seven valued GSK theory, call it $\mathbb{V}^{\mathbf{7A}+}$. Similarly, one can define the GSK theories $\mathbb{V}^{\mathbf{5D}+}$ (in terms of $\mathcal{V}^{\mathbf{3D}+}$ and $\mathcal{K}^+$) and $\mathbb{V}^{\mathbf{7D}+}$ (in terms of $\mathbb{V}^{\mathbf{5D}+}$ and $\mathcal{K}$). In Section 5.6.1, we defined GSK theories of cardinality 5, 6 and 8. We now see that $\mathcal{V}^{\mathbf{3A}+}$ and $\mathcal{V}^{\mathbf{3D}+}$ provide us with the means to define GSK theories of cardinality 7. What about GSK theories of cardinality, say, 9 or 10? Below, we answer this question.

Let's call any expansion exp which is such that both exp and exp^{-1} are contained in $C^{\mathcal{K}^+}_M$, *double* $\mathcal{K}^+$ *closed*. By declaring all stable$_A$ expansions which are double $\mathcal{K}^+$ closed to be open, we obtained $\mathcal{V}^{\mathbf{3A}+}$ from the $\mathcal{K}^+$ closure conditions. Similarly, by declaring all stable$_D$ expansions which are double $\mathcal{K}^+$ closed to be open, we obtained $\mathcal{V}^{\mathbf{3D}+}$ from the $\mathcal{K}^+$ closure conditions. The theories $\mathcal{V}^{\mathbf{3A}+}$ and $\mathcal{V}^{\mathbf{3D}+}$ thus assign, in comparison to $\mathcal{K}^+$, more sentences a "classical assertoric value", i.e., $\mathbf{a}$ or $\mathbf{d}$. Some sentences, such as the Liar, cannot be assigned a classical assertoric value; doing so would result in a semantic valuation which does not respect the identity of truth and which, accordingly, is not a theory of truth in the sense of this paper. An interesting question is whether we can define SK theories which are "more classical" than $\mathcal{V}^{\mathbf{3A}+}$ ($\mathcal{V}^{\mathbf{3D}+}$) in the sense that they assign more sentences a classical assertoric value than $\mathcal{V}^{\mathbf{3A}+}$ ($\mathcal{V}^{\mathbf{3D}+}$). In other words, we may ask whether we can define SK theories which respect $\mathcal{V}^{\mathbf{3A}+}$ ($\mathcal{V}^{\mathbf{3D}+}$). As, according to the $\mathcal{V}^{\mathbf{3A}+}$ and $\mathcal{V}^{\mathbf{3D}+}$ closure conditions, every stable expansion is either closed or open, we have, in order to define such SK theories, turn to the unstable expansions.

Not all unstable expansions are double $\mathcal{K}^+$ closed. An example of an unstable expansion that is contained in $O^{\mathcal{K}^+}_M$ was given above (i.e., expansion (5.24)). However, quite some unstable expansions are double $\mathcal{K}^+$ closed. Below, we discuss a couple of examples. First, consider the expansion associated with an assertion and denial of the Unstabilityteller respectively:

$$A_{T(\mu)}, A_{T(c_0)}, A_{\neg T(c_1)}, D_{T(c_1)}, D_{\neg T(c_2)}, A_{T(c_2)}, \dots \tag{5.28}$$

$$D_{T(\mu)}, D_{T(c_0)}, D_{\neg T(c_1)}, A_{T(c_1)}, A_{\neg T(c_2)}, D_{T(c_2)}, \dots \tag{5.29}$$

In a sense, the Unstabilityteller, $T(\mu)$, is like the Truthteller: there is a (3 valued SK) fixed point in which it is valuated as $\mathbf{a}$ and there is a fixed point in which it is valuated as $\mathbf{d}$. Clearly, in the fixed point where $T(\mu)$ is valuated as $\mathbf{a}$, all the sentences that occur on expansion (5.28) should be valuated in accordance with their sign on (5.28). For instance, $T(c_0)$ should be valuated as $\mathbf{a}$ as its sign on (5.28) is A, whereas $T(c_1)$ should be valuated as $\mathbf{d}$ as its sign on (5.28) is D. Similarly, in the fixed point where $T(\mu)$ is valuated as $\mathbf{d}$, all the sentences that occur on expansion (5.29) should be valuated in accordance with their sign on (5.29). When we want to define a theory that is "more classical" than, say,

$\mathcal{V}^{\mathbf{3A}+}$, we[18] want to declare exactly one of the expansions, (5.28) or (5.29) open. Say that we want to valuate $T(\mu)$ as $\mathbf{a}$. In order to do so, we may declare (5.28) to be open while (5.29) is declared to be closed. However, in order to define a theory of truth via MCG, we need a *systematic* way of judging expansions to be open and closed. The definition of $\mathcal{V}^{\mathbf{3A}+}$ relied on such a systematic way of classifying expansions, as the closure conditions for $\mathcal{V}^{\mathbf{3A}+}$ reveal.

To distinguish unstable expansions in a systematic way, however, is far more tricky. Why do we call (5.28) open? Because it is an unstable expansion whose first term has sign A? Clearly, such a classification will not do, for it results in closure conditions which violate SJC and which do not yield a SK theory, as a moment of reflection on (5.28) testifies. As far as I can see, there simply is no satisfactory systematic way to distinguish expansions (5.28) and (5.29). The lack of such a systematic classification is what prevents me from defining a SK_3 theory which is respected by $\mathcal{V}^{\mathbf{3A}+}$ ($\mathcal{V}^{\mathbf{3D}+}$). However, that is not to say that we cannot define a SK_4 theory which is respected by $\mathcal{V}^{\mathbf{3A}+}$ ($\mathcal{V}^{\mathbf{3D}+}$), as we will now explain.

Although we see no (satisfactory systematic) way to distinguish (5.28) and (5.29), those expansions can (jointly) be distinguished from the following two familiar expansions.

$$A_{\neg T(\lambda)}, D_{T(\lambda)}, D_{\neg T(\lambda)}, A_{T(\lambda)}, A_{\neg T(\lambda)} \dots \tag{5.30}$$

$$D_{\neg T(\lambda)}, A_{T(\lambda)}, A_{\neg T(\lambda)}, D_{T(\lambda)}, D_{\neg T(\lambda)} \dots \tag{5.31}$$

One way to distinguish (5.28) and (5.29) from (5.30) and (5.31) is by observing that the latter two expansions contain a vicious cycle whereas the former two expansions do not. Another way[19] to distinguish them is by observing that the former two expansions are both contained in O^{we}_M whereas the latter two are not. Suppose that we come up with closure conditions which declare more expansions open than the $\mathcal{V}^{\mathbf{3A}+}$ ($\mathcal{V}^{\mathbf{3D}+}$) closure conditions, which satisfy SJC and WRC and according to which both (5.28) and (5.29) are open while both (5.30) and (5.31) are closed. Clearly, such closure conditions will induce a SK_4 theory which respects $\mathcal{V}^{\mathbf{3A}+}$ ($\mathcal{V}^{\mathbf{3D}+}$). Below, we define the theories $\mathcal{V}^{\mathbf{4A}+}$ and $\mathcal{V}^{\mathbf{4D}+}$ according to this rationale, where we distinguish (5.28) and (5.29) from (5.30) and (5.31) by their containment in O^{we}_M. Here are the closure conditions for $\mathcal{V}^{\mathbf{4A}+}$ and $\mathcal{V}^{\mathbf{4D}+}$.

$$O^{\mathbf{4A}+}_M = O^{\mathbf{3A}+}_M \cup (UU_M - ((UU^+_M)^{-1} \cup C^{we}_M))$$
$$O^{\mathbf{4D}+}_M = O^{\mathbf{3D}+}_M \cup (UU_M - ((UU^+_M)^{-1} \cup C^{we}_M))$$

Proposition 5.5 $\mathcal{V}^{\mathbf{4A}+}$ **and** $\mathcal{V}^{\mathbf{4D}+}$ **are** SK_4**.** $\mathcal{V}^{\mathbf{3A}+} < \mathcal{V}^{\mathbf{4A}+}$ **and** $\mathcal{V}^{\mathbf{3D}+} < \mathcal{V}^{\mathbf{4D}+}$
Proof: Similar to the proof of Proposition 5.4. □

5.6.3 More GSK theories

In the previous two subsections, we established the following relations between various SK theories:

[18] Note that an unstable expansion is open according to the $\mathcal{K}^+$ closure conditions just in case it is open according to the $\mathcal{V}^{\mathbf{3A}+}$ ($\mathcal{V}^{\mathbf{3D}+}$) closure conditions.

[19] These really are distinct ways, as is testified by Yablo's paradox. A discussion of that paradox in the present framework is given in Appendix II of this chapter.

$$\mathcal{K} < \mathcal{V}^{\blacklozenge}, \qquad \mathcal{K}^+ \not< \mathcal{V}^{\blacklozenge}$$

$$\mathcal{K} < \mathcal{K}^+ < \mathcal{V}^{we}$$

$$\mathcal{K} < \mathcal{K}^+ < \mathcal{V}^{\mathbf{3A}+} < \mathcal{V}^{\mathbf{4A}+}, \qquad \mathcal{K} < \mathcal{K}^+ < \mathcal{V}^{\mathbf{3D}+} < \mathcal{V}^{\mathbf{4D}+}$$

In this section, we will use the relations in combination with the assertoric transfer theorem to define various GSK theories. We do so in a single sweep via the following table.

SK_3 base	SK_4 base	GSK_5	GSK_6	GSK_7	GSK_8	GSK_{10}
$\mathcal{K}, \mathcal{K}^+$		$\mathbb{V}^{\mathbf{5}+}$				
$\mathcal{K}$	$\mathcal{V}^{\blacklozenge}$		$\mathbb{V}^{\mathbf{6}}$			
$\mathcal{K}, \mathcal{K}^+$	$\mathcal{V}^{we}$		$\mathbb{V}^{\mathbf{6}+}$		$\mathbb{V}^{\mathbf{8}+}$	
$\mathcal{K}, \mathcal{K}^+, \mathcal{V}^{\mathbf{3A}+}$		$\mathbb{V}^{\mathbf{5A}+}$		$\mathbb{V}^{\mathbf{7A}+}$		
$\mathcal{K}, \mathcal{K}^+, \mathcal{V}^{\mathbf{3D}+}$		$\mathbb{V}^{\mathbf{5D}+}$		$\mathbb{V}^{\mathbf{7D}+}$		
$\mathcal{K}, \mathcal{K}^+, \mathcal{V}^{\mathbf{3A}+}$	$\mathcal{V}^{\mathbf{4A}+}$		$\mathbb{V}^{\mathbf{6A}+}$		$\mathbb{V}^{\mathbf{8A}+}$	$\mathbb{V}^{\mathbf{10A}+}$
$\mathcal{K}, \mathcal{K}^+, \mathcal{V}^{\mathbf{3D}+}$	$\mathcal{V}^{\mathbf{4D}+}$		$\mathbb{V}^{\mathbf{6D}+}$		$\mathbb{V}^{\mathbf{8D}+}$	$\mathbb{V}^{\mathbf{10D}+}$

Observe that each row of the table contains at least two SK theories and that these SK theories are written down, from left to right, in accordance with the $<$ relation. For instance, the last row contains 4 SK theories: $\mathcal{K} < \mathcal{K}^+ < \mathcal{V}^{\mathbf{3D}+}$ and $\mathcal{V}^{\mathbf{4D}+}$. The SK theories are used as base theories in order to define, in accordance with the assertoric transfer theorem, the GSK theories on the same row. An example will suffice to clarify the table. Consider the last row of the table, which contains the GSK_{10} theory $\mathbb{V}^{\mathbf{10D}+}$. First, $\mathbb{V}^{\mathbf{6D}+}$ is obtained by applying the assertoric transfer theorem to $\mathcal{V}^{\mathbf{3D}+}$ and $\mathcal{V}^{\mathbf{4D}+}$. As $\mathcal{V}^{\mathbf{3D}+} < \mathbb{V}^{\mathbf{6D}+}$ and as $\mathcal{K}^+ < \mathbb{V}^{\mathbf{3D}+}$, we get $\mathcal{K}^+ < \mathbb{V}^{\mathbf{6D}+}$ and we may apply the assertoric transfer theorem to $\mathcal{K}^+$ and $\mathbb{V}^{\mathbf{6D}+}$ to generate the GSK_8 theory $\mathbb{V}^{\mathbf{8A}+}$. As $\mathcal{K} < \mathbb{V}^{\mathbf{8D}+}$, a last application of the assertoric transfer theorem delivers $\mathbb{V}^{\mathbf{10D}+}$.

Note that the table lacks a GSK_9 theory. The reason for this was explained in the previous subsection: our inability to find a systematic way of distinguishing between unstable expansions like (5.28) and (5.29).

5.7 Concluding remarks

We presented the method of closure games, which is a novel game-theoretic framework for truth and we illustrated in which sense our two stable judgement theorems allow us to study and define 3 and 4 valued SK theories in a uniform manner. By doing so, the method of closure games, sheds new light on SK theories by giving us a better understanding of their interrelatedness than in Kripke's framework. The method of closure games combines ideas from Kripke [33] (Strong Kleene theory of truth), Smullyan [50] (signed tableau calculus) and Fitting [16] (presentation of Kripke's framework using sets of signed sentences rather than models). The main issue of this paper is a presentation of a novel framework for truth and a study of some of its most important (technical) properties. In this last section, we take a more philosophical outlook on the obtained results.

An important distinction between Kripke's framework and ours is the "direction of fit", which is vividly illustrated by contrasting the common definition of $\mathcal{K}_M$ with this paper's definition of $\mathcal{V}^{gr}_M$. Whereas Kripke's "imaginary subject" starts by asserting and denying truth-free sentences and works his way *upwards* until the minimal fixed point is reached, our subject considers the assertion or denial of an arbitrary sentence and sees whether, by following the assertoric rules *downwards*, he can ground his assertoric act in the world. This downwards direction of fit, characteristic for the method of closure games, is also characteristic for the definition of theories of truth by imposing closure conditions on branches, as was detailed in Section 5.5. Another interesting way in which the "downwards methods" of this paper shed new light on Kripke's framework then, is via our proof of the claim that $\mathcal{V}^\bullet = \mathcal{K}^{\mathbf{4}}$. The definition of $\mathcal{V}^\bullet$ was obtained by formalizing a natural assertoric norm—thou shalt respect the world and thou shalt not contradict thyself!—whereas the definition of $\mathcal{K}^{\mathbf{4}}$ is obtained by quantifying over all 3 valued Strong Kleene fixed points. This quantification over all fixed points is what gives $\mathcal{K}^{\mathbf{4}}$ a "modal flavor". It is not so clear that this modal definition of $\mathcal{K}^{\mathbf{4}}$ allows for an intuitively appealing explanation of the claim, say, that the Truthteller is both assertible and deniable. For if, as is quite natural, one conceives of the 3 valued fixed points as a kind of "possible worlds", one is forced to answer the question which of these possible worlds is *actual*. And, naturally, when one asks for the assertoric status of an L_T sentence, one asks for its assertoric status in the actual world. According to $\mathcal{K}^{\mathbf{4}}$, the fact that the Truthteller valuates as $\mathbf{b}$ is explained by the fact that there is a possible world in which it valuates as $\mathbf{a}$ and also, a possible world in which it valuates as $\mathbf{d}$. However, as these possible worlds are not actual, we do not get an intuitive explanation of the fact that, in the actual world, the Truthteller is both assertible and deniable. In sharp contrast, such an intuitive explanation is available on the $\mathcal{V}^\bullet$ picture: by asserting (denying) the Truthteller in the actual world, one takes up assertoric commitments which do not violate the assertoric norm of $\mathcal{V}^\bullet$: thou shalt respect the world and thou shalt not contradict thyself! Hence, we see that our alternative definition of $\mathcal{K}^{\mathbf{4}}$ as $\mathcal{V}^\bullet$ allows for a philosophically more attractive interpretation of the same theory.

The notion of a GSK theory testifies that distinct assertoric norms (closure conditions) can be combined to yield a single compositional theory of truth. An interesting GSK theory is $\mathbb{V}^{\mathbf{8}+}$, an attractive feature of which is that it captures semantic distinctions that are neglected by, for example, $\mathcal{K}$ and $\mathcal{K}^+$ (i.e., these theories equate the Liar and the Truthteller), and that it does so in a compositional way. However, it has to be said that this attractive feature of $\mathbb{V}^{\mathbf{8}+}$ is downplayed by, in my opinion, the lack of an intuitively appealing interpretation. For instance, $\mathbb{V}^{\mathbf{8}+}$ valuates $T(\tau) \wedge \neg T(\tau)$, i.e., the conjunction of the Truthteller with its negation, as $\mathbf{b}_e$. Hence, $T(\tau) \wedge \neg T(\tau)$ is both assertible and deniable. But, one may ask, in which sense is it allowed to assert a contradiction? Technically, the story is clear: It is allowed to assert $T(\tau)$ as doing so does not violate the assertoric norm of $\mathcal{V}^\bullet$. For the same reasons, it is allowed to assert $\neg T(\tau)$. Then, as both its conjuncts are assertible, so is, according to $\mathbb{V}^{\mathbf{8}+}$, $T(\tau) \wedge \neg T(\tau)$. However, in asserting $T(\tau) \wedge \neg T(\tau)$, one violates the assertoric norm of $\mathcal{V}^\bullet$: $\mathcal{V}^\bullet$ valuates $T(\tau) \wedge \neg T(\tau)$ as $\mathbf{d}$. The fact that $\mathbb{V}^{\mathbf{8}+}$ does not share this judgement of $\mathcal{V}^\bullet$ with respect to $T(\tau) \wedge \neg T(\tau)$, is due to its definition in terms of the method of closure games. By asserting $T(\tau) \wedge \neg T(\tau)$, player $\sqcup$ is held responsible, by player $\sqcap$, to assert both its conjuncts. However, in

picking his strategy, player $\sqcap$ must choose one of the conjuncts of $T(\tau) \wedge \neg T(\tau)$. Intuitively, this means that player $\sqcap$ can hold player $\sqcup$ only responsible for one conjunct at a time. This seems to be at odds with the standard interpretation of "asserting a conjunction", according to which, by doing so, one takes up assertoric responsibility for both conjuncts at the same time. So, although $\mathbb{V}^{\mathbf{8}+}$ has, in contrast to $\mathcal{V}^{\bullet}$, a compositional semantics, the price we seem to be paying for its compositionality is an intuitively less satisfying interpretation.

The notion of a GSK theory is, so I claim, a fruitful notion, irrespective of the eventual faith of $\mathbb{V}^{\mathbf{8}+}$. On the one hand, it is technically fruitful, as it aids us in gaining a better understanding of various relations between SK theories. However, Wintein [60] applies the notion of a GSK theory to criticize a proposed desideratum for theories of truth as proposed by Philip Kremer [32]. Let me explain, very briefly, the main point of that paper. In his paper, Kremer provides a theory-relative desideratum for theories of truth. Intuitively, the desideratum, called the *Modified Gupta-Belnap Desideratum* (**MGBD**), says that if there is no vicious reference according to a theory of truth **T** (a formally defined notion that we will not discuss here) then, according to **T**, truth should behave like a classical concept (another formally defined notion that we will not discuss here). Formally:

MGBD If **T** *dictates that there is no vicious reference in ground model* M then **T** *dictates that truth behaves like a classical concept in ground model* M*.*

With respect to the *rationale* of **MGBD**, Kremer cites Gupta [23]:

> For models M belonging to a certain class—a class that we have not formally defined but which in intuitive terms contains models that permit only benign kinds of self-reference—the theory should entail that all Tarski biconditionals are assertible in the model M. (Gupta, [23, p19]

Thus, the proposed rationale for **MGBD** is that it is a theory-relative formalization of an intuitive desideratum that was formulated by Gupta. In [60] an **A**lternative—to **MGBD**—formalization of Gupta's **D**esideratum is proposed:

AD If **T** *dictates that there is no vicious reference in ground model* M then **T** *dictates that all the Tarski biconditionals are strongly assertible in* M*.*

[60] argues that **AD** is preferable over **MGBD** as a desideratum for theories of truth. For one thing, it seems to be superior to **MGBD** in capturing the rationale that is given for that desideratum. In [60], we show that any theory which violates **AD** violates **MGBD**, but also that there are GSK theories of truth (such as $\mathbb{V}^{\mathbf{5}+}$ and $\mathbb{V}^{\mathbf{8}+}$) which violate **MGBD** while they satisfy **AD**. I take it that these results testify that the notion of a GSK theory is a philosophically fruitful notion.

As argued before, the method of closure games gives us a better understanding of SK theories of truth. However, SK theories themselves are often criticized for not defining satisfactory theories of truth. A major criticism is that SK theories lack a "serious conditional"[20]. The material conditional that can

[20] Another is that they suffer from expressive completeness

be defined in SK theories is no serious conditional. For paracomplete theorists such as Field [15], an important criticism of the material conditional that can be defined in $\mathcal{K}$, $\supset$, is that it does not validate the law of identity ($\sigma \supset \sigma$) and neither does it validate the Tarski biconditionals. For paraconsistent theories such as those of Beall [5] and Priest [41], the major criticism of the material biconditional as it can be defined in LP[21] is that it does not validate Modus Ponens. In order to overcome these difficulties and to define a serious conditional, call it $\rightarrow$, all three philosophers help themselves to a semantic approach that consists of a combination of basically two distinct frameworks of truth. Very (very) roughly, connectives other than $\rightarrow$ receive their semantics from a Kripkean (minimal fixed point) construction, while $\rightarrow$ receives its semantics via revision rules as in the Gupta-Belnap framework. An interesting question to ask, and one that I am currently exploring, is whether the method of closure games can provide us with a unified framework in which a serious conditional can be defined. Of course, the requirements on a serious conditional $\rightarrow$ (e.g., it should violate, in light of Curry's paradox, Contraction) ensures that the semantic valuation for a language containing $\rightarrow$ cannot be obtained by (defining the assertoric rules for $\rightarrow$ and by) picking "the right closure conditions". Accordingly, an extension of the framework—by which we determine a valuation of a sentence in virtue of the "power" of player $\sqcup$ to realize expansions of various types—is called for. It is an interesting open question whether the method of closure games allows for a natural extension that can be used to define a serious conditional.

5.8 Appendix I: Proving that $\mathcal{V}^{gr} = \mathcal{K}$

In this section, we prove that, for any ground model M, $\mathcal{V}^{gr}_M = \mathcal{K}_M$, as stated by Proposition 5.1. In order to do so, we first give a constructive definition of $\mathcal{K}_M$, which we adapt from Fitting [16]. Fitting defines the $\mathbf{FP}_M$ set (cf. Definition 5.13) associated with $\mathcal{K}_M$, which we'll denote by $\mathbf{K}_M$ as follows. First, he observes that every set of AD sentences $S \subseteq \mathcal{X}$ has an upwards closure (cf. Definition 5.12) under *the assertoric rules for* L, i.e., under the assertoric rules for L_T as given by Figure 5.4 *minus the assertoric rule which govern the truth predicate*. With S a set of AD sentences, we will use $S^{\Uparrow L}$ to denote its *upwards L closure*, i.e., its upwards closure under the assertoric rules for L. Next, the notion of an upwards L closure is used to define an operator, called Φ, which acts on sets of (AD signed) atomic sentences of L_T (so including atomic sentences of form $T(t)$, where $t \in Con(L_T)$). With S a set of atomic sentences of L_T, $\Phi(S)$ is another such set, where:

$$\Phi(S) = w_M \cup \{X_{T(\overline{\sigma})} \mid X_\sigma \in S^{\Uparrow L}\}$$

With $\rho \in On$, i.e., with ρ an ordinal, $\Phi^\rho(S)$ denotes the ρ fold application of Φ to S, where $\Phi^\rho(S) = \varnothing$ when $\rho = 0$ and where $\Phi^\rho(S) = \bigcup_{\gamma<\rho} \Phi^\gamma(S)$ when ρ is a limit ordinal. Fitting observes that Φ is a monotone operator and that this fact can be exploited to show that, for any set of atomic sentences S, the sequence $\{\Phi^\rho(S)\}_{\rho \in On}$ culminates in a *fixed point*, i.e., there will be some ordinal after which all the terms of the sequence $\{\Phi^\rho(S)\}_{\rho \in On}$ are equal. In particular, there

[21] LP, or logic of paradox, is defined just like $\mathcal{K}$ except that its "middle value" is designated.

will be fixed point for the sequence $\{\Phi^\rho(\varnothing)\}_{\rho \in On}$, i.e., there will be some ordinal, call it Ω, such that $\Phi(\Phi^\Omega(\varnothing)) = \Phi^\Omega(\varnothing)$. This last observation is used by Fitting to define $\mathbf{K}_M$, the $\mathbf{FP}_M$ set associated with $\mathcal{K}_M$, as follows:

$$\mathbf{K}_M = (\Phi^\Omega(\varnothing))^{\Uparrow L} \tag{5.32}$$

In light of the monotonicity of Φ, equation (5.32) can be rewritten as:

$$\mathbf{K}_M = \left(\bigcup_{\rho<\Omega} \Phi^\rho(\varnothing)\right)^{\Uparrow L} = \bigcup_{\rho<\Omega} (\Phi^\rho(\varnothing))^{\Uparrow L} \tag{5.33}$$

The construction of the minimal fixed point as the fixed point of a sequence of sets of signed sentences that is indexed by ordinals closely follows Kripke's [33] original presentation of it. For Kripke, the ordinal levels of the sequence receive an intuitive interpretation in terms of various stages of reflection, by Kripke's "imaginary subject", upon his language. The ordinal levels are also convenient for proving that the minimal fixed point has certain properties, as they allow us to prove claims by transfinite induction. An example of such a claim is given below. However, for the definition of the minimal fixed point, the ordinal levels are, strictly speaking, redundant: the set $\mathbf{K}_M$ can also be defined as the upwards closure (under the assertoric rules of L_T, that is) of the world. Indeed, from the previous definitions of $\mathbf{K}_M$ and the definition of upwards closure, it readily follows that:

$$\mathbf{K}_M = (w_M)^{\Uparrow} \tag{5.34}$$

We now have available the material for proving that $\mathcal{V}_M^{gr} = \mathcal{K}_M$. To do so, it suffices to show that:

$$X_\sigma \in \mathbf{K}_M \Leftrightarrow \exists f \in \mathcal{F} \forall g \in \mathcal{G} : \mathsf{exp}(X_\sigma, f, g) \in G_M^{cor}$$

$\Rightarrow$ Let $X_\sigma \in \mathcal{X}$. In light of equation (5.33), it suffices to show that for any $\rho < \Omega$, $X_\sigma \in (\Phi^\rho(\varnothing))^{\Uparrow L}$ implies that $\exists f \in \mathcal{F} \forall g \in \mathcal{G} : \mathsf{exp}(X_\sigma, f, g) \in G_M^{cor}$. The claim follows by transfinite induction.
$\Leftarrow$ Let $X_\sigma \in \mathcal{X}$ and suppose that for some $f^* \in \mathcal{F}$, we have that $\forall g \in \mathcal{G} : \mathsf{exp}(X_\sigma, f^*, g) \in G_M^{cor}$. So for any $g \in \mathcal{G}$, the expansion $\mathsf{exp}(X_\sigma, f^*, g)$ reaches a signed sentence of $\mathsf{At}_M^*(L)$, called the *ground* of $\mathsf{exp}(X_\sigma, f^*, g)$, which is repeated indefinitely often and which is contained in the world. With $g \in \mathcal{G}$, we write $(Y_\alpha)_g$ to denote the ground of $\mathsf{exp}(X_\sigma, f^*, g)$. The set GR collects all the GRounds and is, given our hypothesis, a subset of the world:

$$GR = \{(Y_\alpha)_g \mid g \in \mathcal{G}\} \subseteq w_M$$

As, relative to f^*, every expansion of X_σ ends up in GR, X_σ is contained in the upwards closure of GR, which is, as $GR \subseteq w_M$, a subset of the upwards closure of w_M which is equal to $\mathbf{K}_M$ (cf. equation (5.34)). I.e.,:

$$X_\sigma \in GR^{\Uparrow} \subseteq (w_M)^{\Uparrow} = \mathbf{K}_M \tag{5.35}$$

This concludes our proof of the claim that, for any ground model M, $\mathcal{V}_M^{gr} = \mathcal{K}_M$.

5.9 Appendix II: Analyzing Yablo's Paradox

A *Yablo sequence* is an infinite sequence of sentences such that, intuitively, each sentence says that all sentences which occur later in the sequence are not true. Each sentence of a Yablo sequence is, just like the Liar, *paradoxical*, meaning that $\mathcal{V}^\bullet$ (cf. Section 5.5.1)) valuates a Yablo sentence as **n**. Although the Yablo sentences share their paradoxical status with the Liar, it is often claimed that, in contrast to the Liar, the Yablo sentences do not exhibit *circular reference* (cf. Yablo [69], who writes that his paradox 'is not in any way circular'). As each sentence of a Yablo sequence talks only about sentences further away in the sequence (and not about it self), the verdict that the Yablo sentences do not exhibit circularity seems in line with our pre-theoretic intuitions. However, the verdict that the Yablo sentences do not exhibit circularity is controversial. Authors (cf. Priest [40]) have disputed this claim and argued, roughly, that the definition of a Yablo sequence relies, just like the definition of the Liar, on a fixed point construction and that *a fortiori*, the Yablo sentences *do* exhibit circularity.

Leitgeb [35] argues that the debate about the circularity of Yablo's paradox is 'substantially flawed'. On the one hand, the two positions in the debate appeal to distinct notions of circularity. On the other hand, Leitgeb argues that all reasonable attempts to spell out these notions in a precise manner seem to surmount in notions that are, in one way or the other, defective. Leitgeb then concludes his paper by asking for the outline of a satisfactory definition of circularity, and he even leaves open the possibility that his question is ill-posed and that talk of circularity is to be banished from science.

> ... either much philosophical work lies ahead of us before the question is finally settled, or that otherwise the question is ill-posed, i.e., that the talk of self-referentiality [and circularity] is to be banished from scientific contexts. (Leitgeb [35, p13])

Urbaniak [52] is an example of work that tries to provide an answer to Leitgeb's question and so is Wintein [61], who gives an account of a notion of circularity by invoking the method of closure games. In [61], I side with Urbaniak, who argues that Leitgeb's *Equivalence Condition*, according to which circularity should be preserved under logical equivalence, is not a reasonable adequacy condition to impose on a definition of circularity. Indeed, the proposed definition of circularity in [61] violates Leitgeb's Equivalence Condition.

In this appendix, however, we will not be concerned with the question as to whether the Yablo sentences are circular or not. What we will do, is discuss how Yablo's paradox is valuated by $\mathcal{V}^\blacklozenge$, i.e., the SK_4 theory that was defined in Section 5.4 via the method of closure games. Our discussion will shed some indirect light on the "circularity debate" however, for consider the closure conditions of $\mathcal{V}^\blacklozenge$:

$$\blacklozenge \textbf{ closure conditions}\text{: } C_M^\blacklozenge = G_M^{inc} \cup U_M^{vic}$$

Indeed, the $\blacklozenge$ closure conditions are spelled out in terms of the notion of a vicious *cycle* (i.e., $U_M^{vic} \subseteq C_M^\blacklozenge$), and so analyzing Yablo's paradox in terms of $\mathcal{V}^\blacklozenge$ will shed some light on relation between circularity and paradoxality, and hence, indirectly, on the"circularity debate". On the other hand, the notion of

circularity that is at work in the ♦ closure conditions pertains to *expansions*, and not, as the (distinct) notions of circularity in the "circularity debate", to sentences. At any rate, we feel that the discussion of Yablo's paradox in terms of $\mathcal{V}^{\blacklozenge}$ is interesting in and of itself. Here we go.

In order to define a Yablo sequence, we let $\{P_n \mid n \in \mathbb{N}\}$, be a sequence of unary predicate symbols. We then define the sequence $\{\sigma_n\}_{n\in\mathbb{N}}$ by letting:

$$\sigma_n := \forall x(P_n(x) \to \neg T(x)) \tag{5.36}$$

The sentences σ_n are formulated using $\to$, i.e., material implication. It will be convenient to explicitly define the assertoric rules for material implication:

†	$A_\dagger$	$D_\dagger$
$\to$	$\dfrac{A_{(\alpha\to\beta)}}{\{D_\alpha, A_\beta\}}\sqcup$	$\dfrac{D_{(\alpha\to\beta)}}{\{A_\alpha, D_\beta\}}\sqcap$

For sake of simplicity, we assume that our ground model M is such that the only way to denote a sentence σ_n is via its quotational name. The elements of $\{\sigma_n\}_{n\in\mathbb{N}}$ will be valuated relative to the world w_M, which satisfies (5.37).

$$A_{P_n(t)} \in w_M \Leftrightarrow t = [\sigma_{n+i}] \text{ for some } i \geq 1 \tag{5.37}$$

In model theoretic terms, (5.37) states that, in w_M, the extension of P_n is equal to $\{\sigma_{n+1}, \sigma_{n+2}, \ldots\}$. Relative to w_M, $\{\sigma_n\}_{n\in\mathbb{N}}$ is a representation of Yablo's sequence: intuitively, σ_n says that every sentence which is *greater than* σ_n is not true. Let f be any strategy for player $\sqcup$ which acts on the denials of the elements of Yablo's sequence as follows:

$$f(D_{\forall x(P_n(x)\to\neg T(x))}) = D_{P_n([\sigma_{n+1}])\to\neg T([\sigma_{n+1}])} \tag{5.38}$$

Let g be any strategy for player $\sqcap$ which acts on the assertions of the elements of Yablo's sequence as follows:

$$g(A_{\forall x(P_n(x)\to\neg T(x))}) = A_{P_n([\sigma_m])\to\neg T([\sigma_m])}, \qquad \text{for some } m > n \tag{5.39}$$

Here is an example of an expansion that is realized with player $\sqcup$ and $\sqcap$ playing strategies respecting (5.38) and (5.39); the full strategies of player $\sqcup$ and $\sqcap$ can be read off from the depicted expansion.

$$A_{\sigma_0}, A_{P_0([\sigma_6])\to\neg T([\sigma_6])}, A_{\neg T([\sigma_6])}, D_{T([\sigma_6])}, D_{\sigma_6}, D_{P_6([\sigma_7])\to\neg T([\sigma_7])}, \ldots$$

$$\ldots D_{\neg T([\sigma_7])}, A_{T([\sigma_7])}, A_{\sigma_7}, A_{P_7([\sigma_8])\to\neg T([\sigma_8])}, A_{\neg T([\sigma_8])}, D_{T([\sigma_8])}, D_{\sigma_8}, \ldots$$

Observe that the expansion is ungrounded but not contained in U_M^{vic} and also, that, relative to f, if player $\sqcap$ picks a strategy which does *not* respect (5.39), then player $\sqcup$ can ensure that a grounded and correct expansion results. It is left to the reader to establish that, with σ_n an arbitrary Yablo sentence, player $\sqcup$ ensures, by playing strategy f, that the expansions of A_{σ_n} and D_{σ_n} are contained in $O_M^{\blacklozenge}$: if player $\sqcap$ deviates from (5.39) a grounded correct expansion will result, while if player $\sqcap$ picks a strategy which respects (5.39), an ungrounded but non-vicious expansion will result. Hence, we have that $\mathcal{V}_M^{\blacklozenge}(\sigma) = \mathbf{b}$.

So, according to $\mathcal{V}^{\blacklozenge}$, the Yablo sentences have a semantic value which is

identical to the value allotted to the Truthteller. In principle, there is nothing (fundamentally) wrong with a theory of truth that allots the Yablo sentences and the Trutheller the same semantic value: the minimal and the maximal intrinsic fixed point, i.e., $\mathcal{K}$ and $\mathcal{K}^+$, do so. However, given the intuitive assertoric interpretation that is associated with the method of closure games, it seems wrong for $\mathcal{V}^{\blacklozenge}$ to do so. It seems that the Yablo sentences are, just like a Liar, *neither* assertible nor deniable, whereas a Truthteller is *both* assertible and deniable. $\mathcal{V}^{\bullet}$ (and $\mathcal{V}^{we}$), however, valuates the Yablo sentences and the Liar as **n**, while it valuates the Trutheller as **b**. More generally, it seems that the norm by which we judge the intuitive (assertoric) plausibility of $\mathcal{V}^{\blacklozenge}$, is captured by $\mathcal{V}^{\bullet}$.

In contrast to $\mathcal{V}^{\blacklozenge}$, however, $\mathcal{V}^{\bullet}$ is not defined in terms of a notion of (expansion) circularity. It is not fair, however, to explain the distinction in the valuation of Yablo sentences between $\mathcal{V}^{\blacklozenge}$ and $\mathcal{V}^{\bullet}$ by appealing to such a notion of circularity. For, note that the closure conditions of $\mathcal{V}^{\Diamond}$ are, modulo the notion of a vicious *cycle*, identical to the $\mathcal{V}^{\blacklozenge}$ conditions, whereas $\mathcal{V}^{\Diamond}$ also valuates a Yablo sentence as **b**. Hence, the valuation distinction between $\mathcal{V}^{\blacklozenge}$ and $\mathcal{V}^{\bullet}$ is better explained in terms of the methods by which these functions are obtained; by playing closure games and by constructing assertoric trees respectively.

Our discussion of Yablo's paradox in terms of $\mathcal{V}^{\blacklozenge}$ suggests that the notion of an expansion is, in a sense, too fine grained to capture the paradoxality of the Yablo sentences. To be sure, $\mathcal{V}^{we}$ is also defined by putting closure conditions on expansions and it does valuate a Yablo sentence as **n**. It does so, however, only because its closure conditions are defined in terms of the closure conditions of $\mathcal{V}^{\blacklozenge}$, which are defined in terms of *branches*, which are *sets* of expansions, not single ones of them. We will now show that, contrary to the suggestion, expansions are not too fine grained to capture the paradoxality of the Yablo sentences. With respect $\mathcal{V}^{\blacklozenge}$'s valuation the Yablo sentences, it is not the $\blacklozenge$ closure conditions that are to be blamed, but rather the incomplete rules of the closure game that defines $\mathcal{V}^{\blacklozenge}$. More concretely, the closure games has no rules that reflect the inferential principles of the "greater than" relation, which seem to be at work in any intuitive reflection on Yablo's paradox. By adding such rules, we will see that $\mathcal{V}^{\blacklozenge}$ *does* valuate the Yablo sentences as **n**.

We will augment the assertoric rules with rules that reflect the fact that, in Yablo's paradox, the *transitive* 'greater-than' relation plays an important role. Here is an intuitive sketch of the way in which the transitivity of the greater-than relation influences the assertoric commitments. Suppose you deny σ_6. That is, suppose that you deny that all elements of $\{\sigma_7, \sigma_8, \ldots\}$ are not true. By denying σ_6, you become committed to assert that at least one element of $\{\sigma_7, \sigma_8, \ldots\}$ is true. And so, if you deny σ_6, you are committed to assert that at least one element of $\{\sigma_5, \sigma_6, \sigma_7, \ldots\}$ is true, i.e., to deny that all elements of $\{\sigma_5, \sigma_6, \sigma_7, \ldots\}$ are not true. Thus, if you deny σ_6 you are, amongst others, committed to deny σ_5. The example easily generalizes, showing that a denial of σ_n commits one to deny σ_m for all $m < n$. Similarly, an assertion of σ_n commits one to assert σ_m for all $m > n$. The assertoric rules for the logical constants that were displayed in the table of Section 2.1. do not allow us to capture this extra-logical reasoning pertaining to the transitive greater-than relation. Therefore, two additional *extra-logical* assertoric rules pertaining to the Yablo sentences are added[22].

[22] I take it that it is more elegant to represent the inferential rules for $>$ (a primitive 'greater than' relation symbol) and then to define a Yablo sequence exploiting $>$. Instead, we presented

$\dagger$	$A_\dagger$	$D_\dagger$
σ_n	$\dfrac{A_{\sigma_n}}{\{A_{\sigma_m} \mid m > n\}}\sqcap$	$\dfrac{D_{\sigma_n}}{\{D_{\sigma_m} \mid m < n\}}\sqcap$

Thus, the (assertoric) Yablo sentences X_{σ_n} have a logical and an extra-logical assertoric rule associated with them. As σ_n is a universally quantified sentence, the type of D_{σ_n} is $\sqcup$. However, according to the extra-logical rule for D_{σ_n}, the type of D_{σ_n} is $\sqcap$. Hence, the question arises which player controls sentences of form D_{σ_n}: is it player $\sqcup$ or player $\sqcap$ who determines which AD sentence is the successor of sentences of form D_{σ_n} in an expansion? In general, a sentence is assertible (deniable) just in case player $\sqcup$ can "live up to all commitments" that arise from an assertion (denial) of that sentence. If a player denies a universal quantification, he is committed to deny an instantiation of the quantified formula. On the other hand, if a player denies a Yablo sentence, he is also committed to deny all "smaller" Yablo sentences. It is in the spirit of the method of closure games to let player $\sqcap$ determine, with respect to D_{σ_n}, which assertoric commitment of player $\sqcup$ is to be considered. Intuitively, when player $\sqcup$ denies σ_n, player $\sqcap$ may either ask player $\sqcup$ which instantiation of $P_n(x) \to \neg T(x)$ he denies or he may conclude that player $\sqcup$ denies, for arbitrary $m < n$, σ_m. To capture these considerations, it is assumed that:

- Sentences of form D_{σ_n} have two types. Thus, both players map D_{σ_n} to one of its immediate AD subsentences. Player $\sqcup$ does so in accordance with the logical rule D_{σ_n}, player $\sqcap$ does so in accordance with the extra-logical rule for D_{σ_n}.
- Player $\sqcap$ determines, for sentences of form D_{σ_n}, the *effective type* of D_{σ_n}. That is, with $f \in \mathcal{F}$ and $g \in \mathcal{G}$, player $\sqcup$ determines whether $f(D_{\sigma_n})$ or $g(D_{\sigma_n})$ succeeds D_{σ_n} in an expansion.
- For sentences of form, A_{σ_n} player $\sqcap$ maps A_{σ_n} on one of its immediate subsentences, either in accordance with the logical or in accordance with the extra-logical rule for A_{σ_n}.

Let us now show that, with the rules of the game just sketched, $\mathcal{V}^\blacklozenge$ valuates the Yablo sentences as **n**. We do so by considering strategies for the players which induce an expansion of A_{σ_n} that is displayed in the table below. In the third column, the type of the sentence under consideration is displayed. For the sentences with two types, both types are displayed and the ineffective type is crossed out; $\sqcup, \not\sqcap$ denotes that the effective type of the sentence under consideration is $\sqcup$.

i	$\mathsf{exp}(i)$	type	i	$\mathsf{exp}(i)$	type
0	A_{σ_n}	$\sqcap$	8	$A_{T([\sigma_m])}$	$\sqcap$
1	$A_{P_n([\sigma_{n+1}])\to\neg T([\sigma_{n+1}])}$	$\sqcup$	9	A_{σ_m}	$\sqcap$
2	$A_{\neg T([\sigma_{n+1}])}$	$\sqcap$	10	$A_{P_n([\sigma_{m+1}])\to\neg T([\sigma_{m+1}])}$	$\sqcup$
3	$D_{T([\sigma_{n+1}])}$	$\sqcap$	11	$A_{\neg T([\sigma_{m+1}])}$	$\sqcap$
4	$D_{\sigma_{n+1}}$	$\sqcap, \not\sqcup$	12	$D_{T([\sigma_{m+1}])}$	$\sqcap$
5	D_{σ_0}	$\not\sqcap, \sqcup$	13	$D_{\sigma_{m+1}}$	$\sqcap, \not\sqcup$
6	$D_{P_0([\sigma_m])\to\neg T([\sigma_m])}$	$\sqcap$	14	D_{σ_m}	$\sqcap, \not\sqcup$
7	$D_{\neg T([\sigma_m])}$	$\sqcap$	15	D_{σ_0}	$\not\sqcap, \sqcup$

the inferential rules to be applicable to the Yablo sentences themselves (as already defined). Doing so is shorter and suffices for our purposes.

Let us illustrate the table by two examples. In step 5, player $\sqcap$ asks player $\sqcup$ which instantiation of $P_0(x) \rightarrow \neg T(x)$ he denies, i.e., player $\sqcap$ determines that the effective type of D_{σ_0} is $\sqcup$. As revealed by the table, it is assumed that player $\sqcup$'s strategy f is such that he answers with σ_m. In step 14, player $\sqcap$ determines that the effective type of D_{σ_m} is $\sqcap$ and, as revealed by the table, he holds player $\sqcup$ responsible for D_{σ_0}. The other steps are explained similarly. Observe that $\{\mathsf{exp}(i) \mid 6 \leq i \leq 15\}$ is a cycle. Moreover, as $\mathsf{exp}(9)$ and $\mathsf{exp}(14)$ testify, $\{\mathsf{exp}(i) \mid 6 \leq i \leq 15\}$ is a *vicious cycle*. Hence, the considered strategies $f \in \mathcal{F}, g \in \mathcal{G}$ and the choice of effective types by player $\sqcap$ results in the depicted expansion of A_{σ_n} that is $\blacklozenge$ closed. It is left to the reader to verify that player $\sqcup$ cannot improve on this outcome by picking another $f \in \mathcal{F}$ than the one which is indicated in the table. Also, it is left to the reader that a similar result holds for D_{σ_n}. Hence, $\mathcal{V}^{\blacklozenge}$ valuates the Yablo sentences as **n**.

So, upon taking care of the assertoric rules pertaining to the "greater than" relation, $\mathcal{V}^{\blacklozenge}$ does acknowledge the paradoxical character of the Yablo sentences. The $\mathcal{V}^{\blacklozenge}$ judgement of those sentences as **n** is explained, on the one hand, by the (use of the) assertoric rules, on the other, by the $\blacklozenge$ closure conditions. Thus, as the $\blacklozenge$ closure conditions are formulated in terms of a notion of (expansion) circularity, $\mathcal{V}^{\blacklozenge}$'s judgement is partly explained by appealing to *a* notion of circularity. However, *the non-compositional* $\mathcal{V}^{\Diamond}$ also—modulo the modified assertoric rules—valuates the Yablo sentences as **n**. As the $\Diamond$ closure conditions are basically the non-circular variant of the $\blacklozenge$ closure conditions, $\mathcal{V}^{\blacklozenge}$'s (indirect) appeal to circularity is, in a sense, not essential for a closure game based explanation that the Yablo sentences are valuated as **n**. Then again, the appeal may be essential for such an explanation via a *compositional* valuation function.

This concludes our discussion of Yablo's paradox in terms of $\mathcal{V}^{\blacklozenge}$. A fuller discussion of these issues and their relation to what we called the "circularity debate" is given in [61].

Chapter 6

Alternative Ways for Truth to Behave when there's no Vicious Reference

6.1 Abstract

In a recent paper, Philip Kremer proposes a formal and theory-relative desideratum for theories of truth that is spelled out in terms of the notion of 'no vicious reference'. Kremer's *Modified Gupta-Belnap Desideratum* (**MGBD**) reads as follows: if theory of truth **T** dictates that there is no vicious reference in ground model M, then **T** should dictate that truth behaves like a classical concept in M. In this paper, we suggest an alternative desideratum (**AD**): if theory of truth **T** dictates that there is no vicious reference in ground model M, then **T** should dictate that all T-sentences are (strongly) assertible in M. We illustrate that **MGBD** and **AD** are not equivalent by means of a *Generalized Strong Kleene theory of truth* and we argue that **AD** is preferable over **MGBD** as a desideratum for theories of truth.

6.2 Introduction

In the paper *How Truth Behaves When There's No Vicious Reference*, [32] is concerned with the behavior of truth under circumstances in which there is *no vicious reference*. Roughly, vicious reference is that type of reference that forces truth—or the truth predicate—to behave in a non-standard manner. The reference involved in a Liar sentence certainly is vicious, while the reference involved in (6.1) certainly is not.

(6.1) consists of 6 words. (6.1)

Kremer argues that our intuitions concerning which sentences exhibit vicious reference and which do not, are (partly) determined by our intuitions concerning which theory of truth is correct. This leads him to suggest that:

> The most general formal articulation of non-vicious reference, we suggest, will be theory-relative. (Kremer [32, p357])

Kremer provides a formal, theory-relative articulation of non-vicious reference and he uses this notion to spell out a formal, theory-relative desideratum for theories of truth. Intuitively, the desideratum, called the *Modified Gupta-Belnap Desideratum* (**MGBD**), says that if there is no vicious reference according to a theory of truth **T**, then, according to **T**, truth should behave like a classical concept. Formally:

MGBD If **T** *dictates that there is no vicious reference in ground model* M, then **T** *dictates that truth behaves like a classical concept in ground model* M.

Kremer compares thirteen theories of truth (ten fixed point theories, three revision theories) in terms of **MGBD**. With respect to the *rationale* of **MGBD**, Kremer cites[1] Gupta [23], who says that:

> For models M belonging to a certain class—a class that we have not formally defined but which in intuitive terms contains models that permit only benign kinds of self-reference—the theory should entail that all Tarski biconditionals are assertible in the model M. (Gupta [23, p19])

Thus, the proposed rationale for **MGBD** is that it is a theory-relative formalization of Gupta's intuitively stated, theory-neutral desideratum—note, Gupta speaks of an *adequacy condition*—for theories of truth. In this paper, we propose an **A**lternative formal and theory-relative translation of Gupta's intuitive **D**esideratum.

AD If **T** *dictates that there is no vicious reference in* M, then **T** *dictates that all the* T*-sentences*[2] *are strongly assertible in* M, where a sentence σ is strongly assertible just in case it is assertible and $\neg\sigma$ is not.

Although any theory which violates **AD** violates **MGBD**, we will see that there are theories of truth which violate **MGBD** and satisfy **AD**. When restricted to the thirteen theories of truth considered by Kremer however, **AD** and **MGBD** are equivalent. The reason of this is that all thirteen theories recognize a *single* semantic value which is allotted to all strongly assertible sentences. This semantic value is, per definition, the same value that is allotted to all *classical* strongly assertible sentences, such as 'snow is white'. Accordingly, with **T** one of theories considered by Kremer, **T** dictates that truth behaves as a classical concept in M just in case **T** dictates that all the T-sentences are strongly assertible in M.

Wintein [63] defined the notion of a *Generalized Strong Kleene theory of truth*, or GSK theory. The distinction between a (three or four valued) Strong Kleene theory of truth and a GSK theory, is that the latter recognizes more than one sense in which a sentence can be strongly assertible. Formally, the semantics of a GSK theory differs from the semantics of a Strong Kleene theory only with respect to negation. Our running example of a GSK theory will be $\mathcal{K}^{\mathbf{5}}$, which has a linear five valued (generalized) Strong Kleene semantics with respect to the lattice:

[1] On page 348 of [32].

[2] A T-sentence, or, in Gupta's words, a Tarski biconditional, is a sentence of form $T(\overline{\sigma}) \leftrightarrow \sigma$, with $\overline{\sigma}$ a closed term which denotes σ.

$$\leq_{\mathbf{5}} := \mathbf{d}_g \leq \mathbf{d}_i \leq \mathbf{e} \leq \mathbf{a}_i \leq \mathbf{a}_g$$

Conjunction and disjunction act as meet and join in $\leq_{\mathbf{5}}$, and universal and existential quantification act as generalized conjunction and disjunction. Negation acts as the identity operation on $\mathbf{e}$ but also, it interchanges $\mathbf{a}_x$ for $\mathbf{d}_x$, where $x \in \{g, i\}$ indicates the *assertoric sense* under consideration: *g*rounded or *i*ntrinsic. The notions of groundedness and intrinsicness reflect that $\mathcal{K}^{\mathbf{5}}$ is defined in terms of Kripke's *Strong Kleene minimal fixed point theory* ($\mathcal{K}$) and his *Strong Kleene maximal intrinsic fixed point theory* ($\mathcal{K}^{+}$).

According to $\mathcal{K}^{\mathbf{5}}$, a sentence is *strongly assertible* just in case its value is contained in $\{\mathbf{a}_g, \mathbf{a}_i\}$, while a sentence is *classical* just in case its value is contained in $\{\mathbf{a}_g, \mathbf{d}_g\}$. Hence, according to $\mathcal{K}^{\mathbf{5}}$ there are non-classical strongly assertible sentences, which explains why $\mathcal{K}^{\mathbf{5}}$ can satisfy **AD** while it violates **MGBD**.

The paper is organized as follows. In Section 6.3, we state some general preliminaries. In Section 6.4, we give the formal definition of **MGBD** and **AD** and show how $\mathcal{K}^{\mathbf{5}}$ testifies that these desiderata are not equivalent. In Section 6.5, we argue that **AD** is preferable over **MGBD** as a desideratum for theories of truth. In Section 5, we are concerned with the assertoric interpretation that we impose on, amongst others, $\mathcal{K}^{\mathbf{5}}$. Section 6.6 concludes.

6.3 Preliminaries

L_T will denote a first order language without function symbols, with *identity* ($\approx$), a *truth predicate* (T) and with a *quotational name* ($[\sigma]$) for each sentence σ of L_T. L will denote the language that is exactly like L_T, except for the fact that it does not contain the truth predicate T. A *ground model* $M = (D, I)$ is an interpretation of L such that $Sen(L_T) \subseteq D$ and such that $I([\sigma]) = \sigma$ for all $\sigma \in Sen(L_T)$. A sentence may be denoted in various ways; $\overline{\sigma}$ will be used to denote any closed term, quotational name or not, which denotes σ in the ground model under consideration. We will make the simplifying assumption that a ground model has, for each of the elements of its domain, a constant symbol which refers to that element. This assumption has the advantage that quantification can be treated substitutionally so that we do not need to be bothered with variable assignments. With respect to $Sen(L_T) \subseteq D$ this assumption is unnecessary, as every sentence contains, per definition, at least one name: its quotational name. However, a sentence may also have a non-quotational name in a ground model, and this feature ensures that a ground model may contain self-referential sentences. Here are some notational conventions that we will respect in this paper concerning the use of some non-quotational names.

Definition 6.1 Some notational conventions
In this paper, the constants λ, τ, η and θ, will be used as follows, where I is an interpretation function.

1. $I(\lambda) = \neg T(\lambda)$. We say that $\neg T(\lambda)$ is a *Liar*.
2. $I(\tau) = T(\tau)$. We say that $T(\tau)$ is a *Truthteller*.
3. $I(\eta) = T(\eta) \vee \neg T(\eta)$. We say that $T(\eta) \vee \neg T(\eta)$ is a *Tautologyteller*.

4. $I(\theta) = T(\theta) \wedge \neg T(\theta)$. We say that $T(\theta) \wedge \neg T(\theta)$ is a *Contradictionteller*.

To be sure, the notational convention does not imply that every ground model contains one of the sentences just defined. However, if we use a sentence which is build with λ, τ, η, or θ, we always presuppose a ground model in which a Liar, Truthteller, Tautologyteller or Contradictionteller occurs. □

As L_T is assumed not to contain function symbols, all the closed terms of L_T are given by its set of constant symbols, which will be denoted by $Con(L_T)$. Observe that $[\forall x T(x)] \approx [\forall x T(x)]$ is guaranteed to be a sentence of L_T. Given a ground model M, $\mathcal{C}_M : Sen(L) \to \{\mathbf{a}, \mathbf{d}\}$ denotes the *classical valuation* of L based on M and is defined as usual[3]. Note that $\mathcal{C}_M([\forall x T(x)] \approx [\forall x T(x)]) = \mathbf{a}$ and $\mathcal{C}_M([\forall x T(x)] \approx [\exists x T(x)]) = \mathbf{d}$ for any ground model M. A *theory of truth* $\mathbf{T}$ takes a ground model as input and outputs a semantic valuation of the sentences of L_T. That is, $\mathbf{T}$ outputs a function $\mathbf{T}_M : Sen(L_T) \to \mathbf{V}$, where $\mathbf{V}$ contains *the semantic values of* $\mathbf{T}$. With $\mathbf{T}$ a theory of truth, $\top_{\mathbf{T}} = \mathbf{T}_M([\forall x T(x)] \approx [\forall x T(x)])$ and $\bot_{\mathbf{T}} = \mathbf{T}_M([\forall x T(x)] \approx [\exists x T(x)])$ are called the *classical top value* and *classical bottom value* of $\mathbf{T}$ respectively. Not any semantic valuation of the sentences of L_T qualifies as the valuation of a theory of truth. In this paper, I assume that in order for $\mathbf{T}$ to qualify as a theory of truth, $\mathbf{T}_M$ should *respect the world*, the *identity of truth*.

Definition 6.2 Theory of truth
Let $\mathbf{T}$ be a valuation method which, given a ground model M, outputs a valuation function $\mathbf{T}_M : Sen(L_T) \to \mathbf{V}$. We say that $\mathbf{T}$ is a theory of truth just in case, for every ground model M, we have that:

$$\forall \sigma \in Sen(L) : \mathcal{C}_M(\sigma) = \mathbf{a} \Leftrightarrow \mathbf{T}_M(\sigma) = \top_{\mathbf{T}}, \;\; \mathcal{C}_M(\sigma) = \mathbf{d} \Leftrightarrow \mathbf{T}_M(\sigma) = \bot_{\mathbf{T}} \quad (6.2)$$

$$\forall \sigma \in Sen(L_T) : \mathbf{T}_M(T(\overline{\sigma})) = \mathbf{T}_M(\sigma) \quad (6.3)$$

That is, $\mathbf{T}_M$ should respect the world (6.2) and the identity of truth [4] (6.3). □

Two interesting three valued theories of truth are Kripke's *Strong Kleene minimal fixed point theory* $\mathcal{K}$, and his *Strong Kleene maximal intrinsic fixed point theory* $\mathcal{K}^+$. In order to define those theories, we let, for every ground model M, $\mathbf{FP}_M$ denote the set of all three valued Strong Kleene fixed point valuations[5] over M. With $V_M, V'_M \in \mathbf{FP}_M$, we let:

$$V_M \leq V'_M \Leftrightarrow \forall \sigma \in Sen(L_T) : V_M(\sigma) = \mathbf{a} \Rightarrow V'_M(\sigma) = \mathbf{a}$$

When $V_M \leq V'_M$ we say that V'_M *respects* V_M. The relation $\leq$ is a partial order on $\mathbf{FP}_M$. The following definitions are all taken from Fitting [16]. We say that V_M is *maximal* just in case for no V'_M we have that $V_M \leq V'_M$, *minimal*

[3] Modulo our use of **a**ssertible and **d**eniable instead of **t**rue and **f**alse, which better fits in with the rest of the paper.

[4] Note that the identity of truth differs from the *intersubstitutability of truth*, according to which $T(\overline{\sigma})$ and σ are interchangeable in every (non opaque) context. In particular, revision theories of truth respects the identity of truth but not its intersubstitutability.

[5] We assume familiarity with the notion of a (three valued) Strong Kleene fixed point valuation over M. To be sure, such a valuation has a Strong Kleene semantics, and it respects the world and the identity of truth. Further, we assume that such a theory valuates sentences of form $T(c)$, where $I(c) \notin Sen(L_T)$ as **d**.

just in case for no V'_M we have that $V'_M \leq V_M$. We say that V_M and V'_M are *compatible* just in case there exists a fixed point[6] V^*_M such that $V_M \leq V^*_M$ and $V'_M \leq V^*_M$. A fixed point V_M is called *intrinsic* just in case it is compatible with every other fixed point. For any ground model M, we let $\mathbf{I}_M$ be the set of all three valued intrinsic fixed points over M. As Kripke [33] shows, $\mathbf{I}_M$ has a maximum element and $\mathbf{FP}_M$ has a minimal element with respect to the relation $\leq$. Using the notions just defined, the definition of $\mathcal{K}$ and $\mathcal{K}^+$ is as follows. Let us remark that we think of the semantic values of $\mathcal{K}$ and $\mathcal{K}^+$ as given by the sets $\{\mathbf{a}, \mathbf{u}, \mathbf{d}\}$ and $\{\mathbf{a}, \mathbf{e}, \mathbf{d}\}$ respectively. When a sentence is valuated as $\mathbf{u}$, we say that that sentence is *ungrounded*, whereas a sentence that is valuated as $\mathbf{e}$ is called *extrinsic*.

Definition 6.3 $\mathcal{K}$ **and** $\mathcal{K}^+$
Let M be an arbitrary ground model. According to the theory $\mathcal{K}^+$, the valuation of L_T in M is given by $\mathcal{K}^+_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{e}, \mathbf{d}\}$, where $\mathcal{K}^+_M$ is (obtained as) the maximum of $\mathbf{I}_M$. According to the theory $\mathcal{K}$, the valuation of L_T in M is given by $\mathcal{K}_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{u}, \mathbf{d}\}$, where $\mathcal{K}_M$ is (obtained as) the minimum of $\mathbf{FP}_M$. □

It is well-known that $\mathcal{K}^+$ respects $\mathcal{K}$, meaning that for every ground model M, we have that $\mathcal{K}_M(\sigma) = \mathbf{a} \Rightarrow \mathcal{K}^+_M(\sigma) = \mathbf{a}$. Not the other way around though: a Tautologyteller is strongly assertible according to $\mathcal{K}^+$ but ungrounded according to $\mathcal{K}$. That is:

$$\mathcal{K}^+_M(T(\eta) \vee \neg T(\eta)) = \mathbf{a}, \qquad \mathcal{K}_M(T(\eta) \vee \neg T(\eta)) = \mathbf{u}$$

Definition 6.4 $\mathcal{K}^{\mathbf{5}}$
The theory $\mathcal{K}^{\mathbf{5}}$ is defined in terms of $\mathcal{K}$ and $\mathcal{K}^+$:

$$\mathcal{K}^{\mathbf{5}}_M(\sigma) = \begin{cases} \mathbf{a}_g, & \mathcal{K}_M(\sigma) = \mathbf{a}; \\ \mathbf{a}_i, & \mathcal{K}_M(\sigma) = \mathbf{u} \text{ and } \mathcal{K}^+_M(\sigma) = \mathbf{a}; \\ \mathbf{e}, & \mathcal{K}^+_M(\sigma) = \mathbf{e}; \\ \mathbf{d}_i, & \mathcal{K}_M(\sigma) = \mathbf{u} \text{ and } \mathcal{K}^+_M(\sigma) = \mathbf{d}; \\ \mathbf{d}_g, & \mathcal{K}_M(\sigma) = \mathbf{d}. \end{cases}$$

□

$\mathcal{K}^{\mathbf{5}}$ is a *Generalized Strong Kleene theory of truth* (*GSK* theory), whose semantics was discussed in the introduction. For the formal definition of the notion of a *GSK* theory—and the proof that $\mathcal{K}^{\mathbf{5}}$ is a *GSK* theory—the reader is referred to Wintein [63].

6.4 The non equivalence of MGBD and AD

6.4.1 Defining MGBD and AD

In this section, we define **MGBD** and **AD** rigorously. That is, we define the following three notions, the first two of which are taken from Kremer [32]:

- **T** dictates that truth behaves like a classical concept in M.

[6] Which we use here as synonymous with 'element of $\mathbf{FP}_M$'.

- **T** dictates that there is no vicious reference in ground model M.

- **T** dictates that all T-sentences are strongly assertible in M.

Definition 6.5 Truth as a classical concept
Let **T** be a theory of truth and M be a ground structure. A sentence σ is *classical*$_{\mathbf{T}}$ *in* M , i.e., classical according to **T** in M, just in case $\mathbf{T}_M(\sigma) \in \{\top_{\mathbf{T}}, \bot_{\mathbf{T}}\}$. A set of of sentences is classical$_{\mathbf{T}}$ in M just in case all its elements are. **T** *dictates that truth behaves like a classical concept in* M just in case $Sen(L_T)$ is classical$_{\mathbf{T}}$ in M. □

Definition 6.6 Strong assertoric pair
Let **T** be a theory of truth, let **V** be the set of semantic values recognized by **T** and let $\{\mathbf{x}, \mathbf{y}\} \subseteq \mathbf{V}$ such that $\mathbf{x} \neq \mathbf{y}$. We say that $\{\mathbf{x}, \mathbf{y}\}$ is a *strong assertoric pair* of **T** just in case we can (linearly) order the elements $\mathbf{x}, \mathbf{y}$ via $<$ such that, for every ground model M and every $\sigma \in Sen(L_T)$:

- Negation acts as a top-bottom swap on $\langle\{\mathbf{x}, \mathbf{y}\}, <\rangle$.

- Conjunction and disjunction acts as meet and join on $\langle\{\mathbf{x}, \mathbf{y}\}, <\rangle$

Given the behavior of $\neg$, $\vee$ and $\wedge$, the *top value* of an assertoric pair $\{\mathbf{x}, \mathbf{y}\}$ of theory **T** may be represented as $\mathbf{T}_M(\sigma \vee \neg\sigma)$, where σ is an arbitrary sentence such that $\mathbf{T}_M(\sigma) \in \{\mathbf{x}, \mathbf{y}\}$. □

Thus, $\{\mathbf{x}, \mathbf{y}\}$ is a strong assertoric pair for a theory of truth **T** just in case, when restricted to $\{\mathbf{x}, \mathbf{y}\}$, negation, conjunction and disjunction allow for a classical algebraic characterization.

Definition 6.7 Strong assertibility of T-sentences
Let **T** be a theory of truth and let M be a ground model. Define $TOP(\mathbf{T})$ as the set of all top values associated with the strong assertoric pairs of **T**[7]. Let $\mathbf{D} \subseteq TOP(\mathbf{T})$ be the members of $TOP(\mathbf{T})$ that are *designated* according to **T**. **T** dictates that all T-sentences are strongly assertible in M just in case:

$$\mathbf{T}_M(T(\overline{\sigma}) \leftrightarrow \sigma) \in \mathbf{D},$$

whenever $\overline{\sigma}$ denotes σ in M. □

Let us make two comments to explain the rationale of Definition 6.7. First, consider Priest's LP interpretation[8] of $\mathcal{K} : Sen(L_T) \to \{\mathbf{a}, \mathbf{u}, \mathbf{d}\}$. According to LP, the designated values of $\mathcal{K}$ are **a** and **u**, implying that—according to LP—all the T-sentences are designated, and so assertible, in every ground model whatsoever. Be that as it may, as the value **u** is not a member of a strong assertoric pair, Definition 6.7 declares that the T-sentences are *not strongly assertible*[9] according to LP. So, by defining the notion of strong assertibility via the notion of a strong assertoric pair, we rule out that both a sentence and its negation can be assertible. Second, Definition 6.7 employs the notion of a *designated* member

[7] Thus, we have that $TOP(\mathcal{K}) = TOP(\mathcal{K}^+) = \mathbf{a}, TOP(\mathcal{K}^{\mathbf{5}}) = \{\mathbf{a}_g, \mathbf{a}_i\}$.

[8] LP abbreviates *Logic of Paradox*.

[9] In every ground model in which there is a sentence that is valuated as **u**, the T-sentence of σ is valuated as **u** and so not strongly assertible.

of $TOP(\mathbf{T})$. The reason for this is that a theory $\mathbf{T}$ assigning σ "the top element of a strong assertoric pair" is by itself no sufficient reason for σ being assertible according to $\mathbf{T}$. For instance, according to (my interpretation of) $\mathcal{K}^{\mathbf{5}}$, sentences which are valuated as either $\mathbf{a}_g$ or $\mathbf{a}_i$ are (strongly) assertible (i.e., $\mathbf{a}_g$ and $\mathbf{a}_i$ are designated), but the formal (algebraic) structure of $\mathcal{K}^{\mathbf{5}}$ does not ensure that the values $\mathbf{a}_g$ and $\mathbf{a}_i$ have to be interpreted as such[10].

What needs to be done in order to complete the definition of **MGBD** and **AD**, is to state Kremer's definition of a theory of truth dictating that there is no vicious reference in M. Before we present this definition, we sketch its rationale. If a set of sentences Y is classical$_{\mathbf{T}}$, then the elements of Y *certainly* do not involve vicious reference. Let $Y \subseteq Sen(L_T)$ be classical$_{\mathbf{T}}$. Then, the *sentential complement* of Y, $\bar{Y} = (Sen(L_T) - Y)$, consists of *potentially* problematic sentences, i.e., of sentences which may (or may not) involve vicious reference. Now if a ground model M can, intuitively, not *discriminate* between the members of $\bar{Y}$, i.e., if M cannot in any way discriminate between potentially problematic sentences, then, according to Kremer's definition, there cannot be vicious reference. M cannot discriminate between members of $X \subseteq Sen(L_T)$ just in case M is X*-neutral*.

Definition 6.8 X**-neutral ground model and clean ground models**
Let $X \subseteq Sen(L_T)$. A ground model $M = (D, I)$ is said to be X*-neutral* just in case:

- For each closed term $t \notin \{[\sigma] \mid \sigma \in Sen(L_T)\}$: $I(t) \notin X$.
- Non-logical predicates do not distinguish between elements of X. That is, with R an n place relation symbol ($R \neq T$) and with $d_1, \ldots, d_n, d_i' \in D$, it holds that if $d_i, d_i' \in X$, then: $(d_1, \ldots, d_i, \ldots d_n) \in I(R) \Leftrightarrow (d_1, \ldots, d_i', \ldots d_n) \in I(R)$.

A $Sen(L_T)$-neutral ground model is called a *clean ground model*. □

Thus, in a clean ground model M, we can refer to sentences only via their quotational names and we cannot discriminate between sentences using any predicate in L. Intuitively, a clean ground model is a ground model with the "least possible amount of vicious reference". Here is Kremer's theory-relative definition of no vicious reference.

Definition 6.9 No vicious reference
Let $\mathbf{T}$ be a theory of truth and let M be a ground model. $\mathbf{T}$ *dictates that there is no vicious reference in* M just in case M is $\bar{Y}$-neutral for some $Y \subseteq Sen(L_T)$ which is classical$_{\mathbf{T}}$ in M. □

Observe that $\varnothing$ is trivially classical$_{\mathbf{T}}$ in M for every theory $\mathbf{T}$. Hence, from definition 6.9 it follows that **MGBD** and **AD** have the following corollaries:

MGBD corollary. If M is a clean ground model, then $\mathbf{T}$ dictates that truth behaves like a classical concept in M.
AD corollary. If M is a clean ground model, then $\mathbf{T}$ dictates that all T-sentences are strongly assertible in M.

[10] I owe this remark to an anonymous referee.

In contrast to **MGBD** and **AD**, their corollaries are partially defined (only for clean ground models) and *theory neutral*, as the notion of a clean ground model is a theory-neutral notion. In a clean ground model, we cannot create Liar sentences, Truthtellers or any other sentences which are, intuitively, problematic. In a clean ground model there cannot, by definition, any kind of vicious reference.

6.4.2 Kremer's results and their consequences for $\mathcal{K}^{\mathbf{5}}$

It can be verified, for arbitrary **T**, whether or not **T** satisfies **MGBD**. As we mentioned, Kremer did so for thirteen theories of truth— ten fixed point theories, three revision theories—in total. The next theorem summarizes Kremer's results.

Theorem 6.1 The results of Kremer
Consider five monotonic valuation schema's: Strong Kleene, Weak Kleene, Supervaluation and two schema's which are called $\sigma 1$ and $\sigma 2$ by Kremer (the details of $\sigma 1$ and $\sigma 2$ do not matter for our purposes). For each of these five schema's, define the associated minimal fixed point and maximal intrinsic fixed point, delivering a total of 10 fixed point theories of truth. With respect to the ten fixed point theories, we have that:

- The maximal intrinsic fixed point theory of any of the five schema's satisfies **MGBD**.

- Only the minimal fixed point theory of $\sigma 2$ satisfies **MGBD**

Further, Kremer considers 3 revision theories of truth:

- $\mathbf{T}^*$, the revision theory of truth based on stability, and $\mathbf{T}^c$, the revision theory of truth based on stability and maximal consistency, satisfy **MGBD**. $\mathbf{T}^{\#}$, the revision theory of truth based on near stability, does not satisfy **MGBD**.

Proof: See [32]. □

We will be concerned[11] mainly with the results pertaining to the Strong Kleene theories $\mathcal{K}$ and $\mathcal{K}^+$. To show that $\mathcal{K}$ violates **MGBD**, it suffices to consider a clean ground model M_0. With $\mathsf{LEM} := \forall x(T(x) \vee \neg T(x))$, we have that $\mathcal{K}_{M_0}(\mathsf{LEM}) = \mathbf{u}$, and so the result readily follows: there is no vicious reference in M_0 according to $\mathcal{K}$ and yet truth does not behave as a classical concept in M_0 according to $\mathcal{K}$. Similarly, the fact that $\mathcal{K}$ violates **AD** follows from the observation that $\mathcal{K}_{M_0}(T(\lceil\mathsf{LEM}\rceil) \leftrightarrow \mathsf{LEM}) = \mathbf{u}$. Further, we have that:

$$\mathcal{K}^{\mathbf{5}}_{M_0}(\mathsf{LEM}) = \mathcal{K}^{\mathbf{5}}_{M_0}(T(\lceil\mathsf{LEM}\rceil) \leftrightarrow \mathsf{LEM})) = \mathbf{a}_i$$

[11] As argued by Kremer, his obtained results put doubt on Gupta and Belnap's claim that revision theories of truth have, as a distinctive general advantage over fixed point theories, their '...consequence that truth behaves like an ordinary classical concept under certain conditions—conditions that roughly can be characterized as those in which there is no vicious reference in the language.' ([24, p201]). As the claim of Gupta and Belnap is cast in terms of the intuitive theory-neutral notion of no vicious reference, Kremer's results cannot be said to falsify their claim.

Thus, $\mathcal{K}^{\mathbf{5}}$ does not dictate that truth behaves like a classical concept in M_0, the reason being that there are sentences, such as LEM, which are not classical$_{\mathcal{K}^{\mathbf{5}}}$ in M_0, as they are not valuated as $\mathbf{a}_g$ or $\mathbf{d}_g$. As there is no vicious reference in M_0 according to $\mathcal{K}^{\mathbf{5}}$, it follows that $\mathcal{K}^{\mathbf{5}}$ violates **MGBD**. On the other hand, from the definition of $\mathcal{K}^{\mathbf{5}}$ and the fact that $\mathcal{K}^{+}$ satisfies **MGBD**, it immediately follows that:

$\mathcal{K}^{\mathbf{5}}$ dictates that there is no vicious reference in $M \Rightarrow$
$\forall \sigma \in Sen(L_T) : \mathcal{K}^{\mathbf{5}}_M(\sigma) \in \{\mathbf{a}_g, \mathbf{a}_i, \mathbf{d}_g, \mathbf{d}_i\}$

Moreover, from the compositionality of $\mathcal{K}^{\mathbf{5}}$, it follows that:

$\mathcal{K}^{\mathbf{5}}$ dictates that there is no vicious reference in $M \Rightarrow$
$\mathcal{K}^{\mathbf{5}}_M$ is *linear* (generalized) Strong Kleene w.r.t. $\mathbf{d}_g \leq \mathbf{d}_i \leq \mathbf{a}_i \leq \mathbf{a}_g$.

From the behavior of $\mathcal{K}^{\mathbf{5}}$ with respect to $\mathbf{d}_g \leq \mathbf{d}_i \leq \mathbf{a}_i \leq \mathbf{a}_g$, it follows that a T-sentence of σ is valuated as $\mathbf{a}_g$ when σ is valuated as $\mathbf{a}_g$ or $\mathbf{d}_g$, and as $\mathbf{a}_i$ when σ is valuated as $\mathbf{a}_i$ or $\mathbf{d}_i$. Accordingly, we get that:

$\mathcal{K}^{\mathbf{5}}$ dictates that there is no vicious reference in $M \Rightarrow$
$\forall \sigma \in Sen(L_T) : \mathcal{K}^{\mathbf{5}}_M(T(\overline{\sigma}) \leftrightarrow \sigma) \in \{\mathbf{a}_g, \mathbf{a}_i\}$

Thus, if $\mathcal{K}^{\mathbf{5}}$ dictates that there is no vicious reference in M, all the T-sentences will be strongly assertible in M. To sum up, we get:

Theorem 6.2 $\mathcal{K}^{\mathbf{5}}$ **violates MGBD and satisfies AD**
Proof: Given above. □

6.4.3 The intrinsic hedge

$\mathcal{K}^{\mathbf{5}}$ satisfies **AD** due to the protection, against a violation of **AD**, that it obtains from $\mathcal{K}^{+}$. There are more theories than $\mathcal{K}^{\mathbf{5}}$ which may obtain such "intrinsic protection". Kremer considers five distinct monotonic valuation schemes in total, for each of which he defines the minimal fixed point theory and the maximal intrinsic fixed point theory. All five considered maximal intrinsic fixed points satisfy[12] **MGBD**. Accordingly, all five minimal fixed point theories have the possibility to hedge against a violation of **AD** by buying protection from their intrinsic cousins by defining a five valued theory in a manner similar[13] to the definition of $\mathcal{K}^{\mathbf{5}}$. However, we will only be concerned with the Strong Kleene version of the intrinsic hedge.

[12] Kremer [31] gives an elegant proof (Theorem 4.21, 2.iv) which establishes a more general result. Given some very weak conditions on a partial function $\mathcal{F}$, defined on *hypotheses*—potential significations of the truth predicate—the maximal intrinsic fixed point theory associated with $\mathcal{F}$ satisfies **MGBD**. As the partial functions associated with each of the five schema's considered by Kremer satisfy the mentioned conditions, the associated maximal intrinsic fixed point theories all satisfy **MGBD**.

[13] In a similar vein, it may also be possible for $\mathbf{T}^{\#}$, the revision theory of truth which violates **MGBD** and **AD**, to hedge against a violation of **AD** by buying protection from $\mathbf{T}^{*}$, which satisfies **MGBD**. However, as we are not interested in "saving $\mathbf{T}^{\#}$" from the violation of any desideratum in the first place, we do not touch these matters.

6.5 AD or MGBD as a desideratum for theories of truth?

6.5.1 The fundamental intuition about truth

As mentioned in the introduction, Kremer cites Gupta [23] with respect to the rationale of **MGBD**. The following quote of Gupta directly precedes the quote that was given in the introduction.

> We conclude, then, that in a variety of circumstances we can consistently maintain the fundamental intuition. I suggest that it is a reasonable adequacy condition on any theory that purports to explain the meaning of 'true' that under such circumstances it preserve the fundamental intuition—or at least the intuition should be preserved if it does not come into conflict with some other intuitions that are of equal or greater importance.
>
> (Gupta [23, p19])

When we follow the above suggestion of Gupta, and also Kremer's suggestion that vicious reference is a theory-relative phenomenon, we get the following (abstract and informal) desideratum for theories of truth, that is cast in terms of the "fundamental intuition about truth". Here is the *Fundamental Intuition Desideratum* (**FID**):

FID If **T** dictates that there is no vicious reference in M, then **T** preserves the *Fundamental Intuition about truth* (FI).

One way[14] to understand the relation between **AD** and **MGBD** is via **FID**. For, we may understand both **AD** and **MGBD** as agreeing in their endorsement of **FID**, while they disagree over how **FID** should be spelled out concretely. For, **AD** and **MGBD** rely on two different specifications of the FI, i.e., on FI_1 and FI_2 respectively:

FI_1: All Tarski biconditionals should be assertible.

FI_2: Truth should behave like a classical concept.

When we understand the relation between **AD** and **MGBD** via **FID**, we are forced to argue, in order to claim that **AD** is preferable over **MGBD**, that FI_1 is a better precisification of FI than FI_2. Here are some arguments in favor of FI_1 over FI_2.

1) Specific and non-specific intuitions about truth. In contrast to FI_1, FI_2 is not an intuition that specifically pertains to truth. Rather, FI_2 seems to be an instantiation—with the concept of truth—that we have with respect to any concept X whatsoever:

- X should behave like a classical concept.

[14] See Section 4.2 for another way.

Being an instantiation of such a general schema, it seems awkward to speak of FI_2 as the fundamental intuition *about truth*. Arguably, fundamental intuitions about any concept X whatsoever should be spelled out in terms of characteristic features of X.

2) Non-sentential objects and category mistakes. Truth is *appropriately* ascribed, one may say[15], to sentences. Truth ascriptions to non-sentential objects are, in an important sense, inappropriate. For instance, to say that 'snow is true' is to say something inappropriate in a sense in which saying that "snow is black' is true' is not. Typically, one abstracts away from the difference in inappropriateness alluded to: one simply holds that 'snow is true' and "snow is black' is true' are on a par in the sense that they are both false[16] This is a convenient abstraction on which, typically, not much hinges. However, suppose, for the sake of argument, that philosopher P holds that to ascribe truth to snow is make a category mistake and that, *a fortiori*, the sentence 'snow is true' is, say, neither true nor false. Yet any clean ground model whose domain contains non-sentential objects will contain sentences that involve category mistakes of the 'snow is true' type. As these sentences are, according to P, neither true nor false, truth will not behave like a classical concept in such clean ground models. Hence, any theory of truth that respects P's view on category mistakes violates **MGBD**. In contrast, as FI_1 is formulated in terms of the *T-sentences*, there may be theories of truth that respects both P's view on category mistakes and **AD**. In light of P's situation, a proponent of **MGBD** has to argue that P's view on category mistakes not only conflicts with FI_2 (which it does), but that this is *how it should be*. In concreto, he has to argue that it belongs to the fundamental intuition about truth that 'snow is true' is false (or true, but that's absurd). However, I do not see how P's view on category mistakes conflicts with any "fundamental intuition about truth" whatsoever, and I take this as evidence in favor of **AD** over **MGBD**.

To put some more flesh on the bones, consider another philosopher, P' which has the following view on sentences such as 'snow is true'. According to P', it is outrageous to assert 'snow is true'. With respect to denying 'snow is true', however, he is less explicit. He doesn't think that denying 'snow is true' is outrageous, but he doesn't want to deny it in the same sense as he denies 'snow is black'. P' can help himself to a theory of truth which respects his intuitions and **AD** (but violates **MGBD**) as follows. Remember that the set $\mathbf{FP}_M$ of all Strong Kleene fixed point valuations over M was defined as follows (cf. footnote 5): $V_M \in \mathbf{FP}_M$ just in case,

1. V_M respects the world.
2. V_M respects the identity of truth.
3. V_M has a Strong Kleene semantics.
4. $V_M(T(c)) = \mathbf{d}$ whenever $I(c) \notin Sen(L_T)$

[15] We do not enter the discussion of whether it is more appropriate to ascribe truth to, say, propositions. Nothing substantial will hinge on the assumption that the truth-bearers are sentences.

[16] Note that Kremer's results (cf. Theorem 6.1) depend on the assumption that each of the thirteen theories of truth valuates truth ascriptions to non-sentential objects as (in our terms) **a** or **d** (Kremer naturally assumes the latter).

In light of his intuitions, P' defines the set $\mathbf{FP}^{\star}_{M}$, which is defined by keeping conditions 1,2 and 3 as they are and by trading in 4 for $4'$, where:

$4'$ $V_M(T(c)) \neq \mathbf{a}$ whenever $I(c) \notin Sen(L_T)$

Condition $4'$ reflects that P' thinks that it is outrageous to assert 'snow is true'. P' constructs the minimal fixed point and the maximal intrinsic fixed point over $\mathbf{FP}^{\star}_{M}$ and combines them into a five valued GSK theory, call it $\mathcal{K}^{\mathbf{5}\star}$, as before. Observe that, according to $\mathcal{K}^{\mathbf{5}\star}$, sentences as 'snow is true' will be valuated as $\mathbf{d}_i$. Such sentences are deniable, but not in the same sense as 'snow is black' (which is valuated as $\mathbf{d}_g$). By a similar argument as before, $\mathcal{K}^{\mathbf{5}\star}$ respects **AD** but violates **MGBD**.

To sum up, in order to satisfy **MGBD**, a theory of truth has to valuate non sentential truth ascriptions classically, while this is not so for **AD**. There are certain views on category mistakes which do not seem to conflict with any fundamental intuition about truth whatsoever and according to which non sentential truth ascriptions should not be valuated classically. In light of such views, **AD** seems preferable over **MGBD**.

3) The central place of FI_1 in the literature on truth. We're approaching the question as to whether **AD** is preferable over **MGBD** as a desideratum for theories of truth via the question as to whether FI_1 or FI_2 is a better precisification of FI. As such, the exhaustive appeal in the literature (in some way or other) to FI_1 as the fundamental intuition about truth (and not to FI_2) constitutes a clear dialectical advantage for favoring FI_1 or FI_2. I will not bother the reader with an exhaustive list of quotes from authors such as Aristotle, Tarski, Horwich, Gupta who all appeal, in some way or other, to FI_1 as the fundamental intuition about truth. Our results established that FI_1 and FI_2 result in non-equivalent desiderata for theory of truth. Given the central place of FI_1 in the literature on truth, we feel that the burden of proof is on the side of those who claim that FI_2 is a better precisification of FI than FI_1.

6.5.2 Reasoning classically

Now, a proponent of **MGBD** may argue that the rationale of **MGBD** should not be understood in terms of **FID**. That is, **MGBD** should not be understood as stating conditions under which the fundamental intuition about truth (which, so he may admit, is FI_1) should be respected. Rather, there is an independent rationale for **MGBD**. For instance, when Gupta and Belnap speak about the Gupta-Belnap Desideratum, they do not refer to fundamental intuitions about truth at all:

> An important feature of the revision theory, and one that prompted our interest in it, is its consequence that truth behaves like an ordinary classical concept under certain conditions—conditions that can roughly be characterized as those in which there is no vicious reference in the language. (Gupta & Belnap [24, p201])

This quote occurs in the beginning of chapter 6 of *The Revision Theory of Truth* and a large part of that chapter is devoted to finding circumstances under which truth behaves like a classical concept. In that chapter, Gupta and

Belnap do not relate the notion of truth's classical behavior to the assertibility of the T-sentences. This suggests that truth's classical behavior under favorable circumstances may be desirable for reasons that are not spelled out in terms of **FID**. Arguably, this is what Gupta and Belnap have in mind:

> [The Gupta-Belnap Desideratum] captures the intuition that if there is no vicious reference in the language then our ordinary ways of working with the concept of truth are unproblematic and, consequently, the interpretation of truth should be classical.
>
> (Gupta & Belnap [24, p112])

Thus, the rationale of the Gupta-Belnap Desideratum (which, we take it, carries over to **MGBD**) is as follows:

1. If there's no vicious reference one should be able to reason with truth in accordance with our ordinary ways.

While, in order for this rationale to be realized, Gupta and Belnap take it that:

2. Consequently, when there's no vicious reference, truth should behave like a classical concept.

Gupta and Belnap argue that their revision theories satisfy their desideratum, for:

> We have seen that in models that permit only certain restricted kinds of self-reference, the revision process yields a classical extension as the signification of truth. In these models, all our unreflective intuitions are preserved: Classical principles of reasoning hold; so do all the Tarski biconditionals [...]; so also do the semantic principles (e.g., the principle that a conjunction is true iff its conjuncts are true). (Gupta & Belnap [24, p219])

Now, observe that the definition of strong assertibility (in terms of the notion of a strong assertoric pair) precisely ensures that, whenever all sentences of L_T are strongly assertible or deniable in ground model M (and so all the T-sentences are strongly assertible), the classical principles of reasoning hold. Indeed, $\mathcal{K}^{\mathbf{5}}$ testifies that, for the classical principles of reasoning to hold, it is not required that truth behaves like a classical concept. Hence, a proponent of **AD** may accept the rationale of **MGBD** (claim 1) while we have shown that, in order for this rationale to be realized, it is not necessary that truth behaves like a classical concept (claim 2). I take it that this establishes that, in the debate between an **AD** and **MGBD** proponent, the burden of proof shifts to the latter one; (s)he owes us an argument why $\mathcal{K}^{\mathbf{5}}$, which validates the classical principles of reasoning when there's no vicious reference, is an "undesirable" theory of truth due to its violation of **MGBD**.

6.6 On the interpretation of $\mathcal{K}^{\mathbf{5}}$

6.6.1 An objection

In this section, we will make a couple of remarks on the assertoric interpretation of ($\mathcal{K}, \mathcal{K}^{+}$ and) $\mathcal{K}^{\mathbf{5}}$. The assertoric interpretation that we imposed on the

(Generalized) Strong Kleene theories of truth is reflected by our use of **a** and **d**, possibly with subscripts. In this section, we will also consider alternative interpretations of such theories, and it will be convenient to work with a more neutral notation for their range: we let $V_M : Sen(L_T) \to \{0, \frac{1}{2}, 1\}$ be a Strong Kleene fixed point valuation over ground model M. V_M^{min} and V_M^+ will denote the minimal and maximal intrinsic fixed point in this notation.

With respect to the interpretation of V_M^{min}, both Field [15] and Kremer [30] have convincingly argued[17] that the external notion of having semantic value 1 cannot be equated with the internal notion of truth in the interpreted language (L_T, V_M^{min}) that is expressed by the truth predicate of L_T. Here is why we can't do that. Consider a Liar sentence $\neg T(\lambda)$. Now, $V_M^{min}(\neg T(\lambda)) = \frac{1}{2} \neq 1$. So, *if we equate having semantic value 1 with truth*, it follows that the Liar is not true. Expressing this claim in L_T, we thus need to have that $V_M^{min}(\neg T([\neg T(\lambda)])) = 1$. But in fact, we have that $V_M^{min}(\neg T([\neg T(\lambda)])) = \frac{1}{2}$. As Field rightly remarks:

> [...] these decidedly odd features are not consequences of Kripke's construction. They result, rather, from the identification of truth with having semantic value 1. (Field [15, p69])

The assertoric interpretation of V^{min} as $\mathcal{K}$, does not equate having semantic value 1 with truth and so avoids the "decidedly odd features" that Field is referring to. We say that $\mathcal{K}_M(\neg T(\lambda)) = \mathbf{u}$, i.e., that the Liar is *ungrounded*, but the ungroundedness of a sentence is not related to its *truth*-value.

From the perspective of $\mathcal{K}$, the ungroundedness of a sentence can be taken to indicate that it is neither assertible nor deniable[18]. Hence, $\mathcal{K}_M(\sigma) = \mathbf{u}$ can be paraphrased as 'σ is neither assertible nor deniable according to (the assertoric norm that is associated with) $\mathcal{K}_M$'. Similarly, *from the perspective of* $\mathcal{K}_M^+$, $\mathcal{K}_M^+(\sigma) = \mathbf{e}$ can be paraphrased as 'σ is neither assertible nor deniable according to (the assertoric norm that is associated with) $\mathcal{K}_M^+$'. The relation between $\mathcal{K}_M$ and $\mathcal{K}_M^+$ indicates that these theories are associated with distinct assertoric norms and that the $\mathcal{K}_M$ norm is stricter than the $\mathcal{K}_M^+$ norm. According to $\mathcal{K}^{\mathbf{5}}$, the $\mathcal{K}_M$ norm is too strict: the ungroundedness of a sentence is not a reason to render it neither assertible nor deniable. On the other hand, $\mathcal{K}^{\mathbf{5}}$ acknowledges that assertible (deniable) ungrounded sentences cannot be asserted (denied) in the same sense as grounded ones.

Although our assertoric interpretation of (Generalized) Strong Kleene theories of truth does not have the "decidedly odd features" that are associated with interpretations that equate having semantic value 1 with truth, there is a worry that the assertoric interpretation faces a problem of its own, which can be stated as the following objection:

Obj(ection): Assertibility does not behave in line with the sketched assertoric interpretation of a Strong Kleene theory of truth. In particular, there are *grounded* sentences which are neither assertible nor deniable yet according to $\mathcal{K}_M$, such sentences will be valuated as **a** or **d**.

To illustrate **Obj**, consider the following sentence:

Ceasar had exactly 12 hairs on his big toe as he crossed the Rubicon. (6.4)

[17] In fact, their argument pertains to any Strong Kleene fixed point valuation.

[18] We take it that a sentence is deniable just in case its negation is assertible.

Clearly, (6.4) is a grounded sentence, and so $\mathcal{K}_M$ will valuate it either as **a** or **d**. Yet it seems absurd to claim that (6.4) is either assertible or deniable, as nobody knows (6.4) or its negation[19].

Below, I will formulate two distinct replies to this objection. The first reply observes that **Obj** only goes through on certain accounts of assertion which have been disputed by, amongst others, [55]. Further, we observe that on the account of assertion as proposed by Weiner, we can make good sense of $\mathcal{K}_M$'s assertoric interpretation.

In contrast, our second reply accepts the knowledge norm of assertion. We take it that $\mathcal{K}_M$ delivers the correct assertoric verdict with respect to grounded sentences only for ground models in which sentences like (6.4) do not occur. Then, we show how $\mathcal{K}_M$ can be represented via the *method of closure games*, a game theoretic method that can be used to define theories of truth. We show how $\mathcal{K}_M$'s representation via the method of closure games can be generalized to deliver a theory of truth that also gives the right verdicts with respect to ground models in which sentences like (6.4) *do* occur. The theory of truth thus obtained reveals that the main argument of this paper —pointing out the distinctions between **AD** and **MGBD** and arguing that **AD** is preferable—is not threatened by **Obj**.

6.6.2 The knowledge norm, Weiner's norm and the norms of $\mathcal{K}$ and $\mathcal{K}^+$

Obj relies on a specific assertoric norm: one should assert σ only if one knows σ. Plausible as the knowledge norm of assertion may seem, it is far from indisputable. For instance, Weiner [55] argues that:

> [...] it is possible to explain the cases that motivate the knowledge account of assertion by postulating a general norm that assertions would be true, combined with conversational norms that govern all speech acts. A theory on which proper assertions must be true explains the data better than a theory on which assertions must be known to be true. (Weiner [55, p227])

It is outside the scope of this paper to assess Weiner's theory in any detail. I will just observe that, if Weiner's account of assertion is correct, **Obj** does not go through. Consider sentence (6.4). Depending on the ground model M, (6.4) is either true or false. Suppose that M is such that (6.4) is true. Then, according to Weiner's *general norm of assertion*, (6.4) is assertible, which is in accordance with the verdict of $\mathcal{K}_M$ with respect to (6.4). At the same time, an assertion of (6.4) is improper on Weiner's theory, but the *improperness is not to be explained by the general norm of assertion*. Rather, it is to be explained by violations of "conversational norms that govern all speech acts", which are spelled out by Weiner as Gricean maxims. Building on Weiner's account of assertion, we can think of $\mathcal{K}_M$ as expressing verdicts that derive from a general norm of assertion (a sentence being assertible when it is *grounded* and true), while the improperness of assertions (or denials) of sentences like (6.4) is to be explained via conversational norms that govern all speech acts.

[19] For our purposes, we may also say that (6.4) is neither assertible nor deniable as nobody has enough evidence to justify an assertion or denial of (6.4).

The assertoric norm of $\mathcal{K}^+_M$ is more liberal; a sentence is assertible according to $\mathcal{K}^+_M$ iff it is assertible according to $\mathcal{K}_M$ or if (it is ungrounded and) its assertion can be justified on *intrinsic grounds*. An assertion of σ can be justified on intrinsic grounds just in case:

1. By denying σ one becomes committed to contradict oneself.
2. By asserting σ one does not become committed (to contradict oneself nor to) an *arbitrary* assertion or denial.

One can become committed to an assertion of σ in two ways: via an outright assertion of σ or indirectly, as when one becomes committed to an assertion of α by assertion $\alpha \wedge \beta$. As explained in Wintein [63], we take it that the transmission of assertoric commitments is governed by the assertoric rules for L_T, amongst which are:

T	$\dfrac{A_{T(\overline{\sigma})}}{A_\sigma}$	$\dfrac{D_{T(\overline{\sigma})}}{D_\sigma}$

$\wedge$	$\dfrac{A_{(\alpha\wedge\beta)}}{A_\alpha, A_\beta}$	$\dfrac{D_{(\alpha\wedge\beta)}}{D_\alpha \mid D_\beta}$

$\neg$	$\dfrac{A_{\neg\sigma}}{D_\sigma}$	$\dfrac{D_{\neg\sigma}}{A_\sigma}$

To illustrate these two claims, let $\neg T(\lambda)$ be a Liar, $T(\tau)$ be a Truthteller and let $T(\eta) \vee \neg T(\eta)$ be a Tautologyteller (cf. Definition 6.1). Consider the sentence $\neg T(\lambda) \vee T(\tau)$ first. By denying $\neg T(\lambda) \vee T(\tau)$, one becomes committed to deny $\neg T(\lambda)$ and to deny $T(\tau)$ (rule $D_\vee$). The commitment to deny $\neg T(\lambda)$ results in a commitment to assert $T(\lambda)$ (rule $D_\neg$) which results in a commitment to assert $\neg T(\lambda)$ (rule A_T). Hence, by denying $\neg T(\lambda) \vee T(\tau)$ one becomes committed to assert and deny $\neg T(\lambda)$, i.e., one becomes committed to contradict oneself. So, $\neg T(\lambda) \vee T(\tau)$ satisfies the first condition of being assertible on intrinsic grounds. Not the second condition though. For, by asserting $\neg T(\lambda) \vee T(\tau)$ one becomes committed to assert $\neg T(\lambda)$ or to assert $T(\tau)$. Asserting $\neg T(\lambda)$ leads to a self-contradiction, but to assert the Truthteller $T(\tau)$ is to make a completely arbitrary assertoric move: (according to $\mathcal{K}^+_M$) there is nothing which favors an assertion of a Truthteller over a denial. Hence, condition 2 is not fulfilled and $\neg T(\lambda) \vee T(\tau)$ is not assertible on intrinsic grounds. In contrast, the Tautologyteller $T(\eta) \vee \neg T(\eta)$ is assertible on intrinsic grounds, as the reader may verify. As we explained above, $\mathcal{K}^{\mathbf{5}}$ combines the assertoric norms of $\mathcal{K}$ and $\mathcal{K}^+$.

6.6.3 (Non-) Omniscient Agent Models

Here is another reply to **Obj** which accepts the knowledge norm of assertion. Remember that we took a ground model M to induce a classical valuation $\mathcal{C}_M : Sen(L) \to \{\mathbf{a}, \mathbf{d}\}$ and that a theory of truth must respect this classical valuation of L. Hence, we assumed that, relative to a ground model M, all L sentences are either **a**ssertible or **d**eniable. As we accept the knowledge norm of assertion, we are committed to hold that the ground models considered thus far are such that any L sentence σ or its negation is known. Thus, sentences like (6.4) simply do not occur in those ground models, which we will call *Omniscient Agent models* (OA models) from now on. With respect to OA models, $\mathcal{K}_M$'s verdicts are in line with the knowledge norm of assertion. In this section, we will show how to generalize the main features of $\mathcal{K}_M$ so that it becomes applicable to *Non-Omniscient Agent models* (NOA models) as well. In order to do

so, we rely on a representation of $\mathcal{K}_M$ via the method of closure games, a game theoretic valuation method for theories of truth that is developed in Wintein [63]. First, we represent the main features of the method of closure games, after which we show how it induces $\mathcal{K}_M$ for an OA model M. Then, we show that $\mathcal{K}_M$'s representation via the method of closure games can be generalized so that it becomes applicable to NOA models as well.

Inducing theories of truth via the Method of Closure games.
The *Method of Closure Games* (MCG) is a framework for truth, which defines a theory of truth upon the specification of a *closure condition*. Intuitively, a closure condition specifies, for each ground model M (here, thought of as a OA model), the conditions under which sentences are assertible and deniable. The central notion in MCG is that of an *expansion*: a closure condition assigns, to each ground model M, a bipartition of the set of all expansions into the sets of *open* and *closed* expansions in M. The notion of an expansion is defined in terms of the *assertoric rules of* L_T, which are basically the rules of a signed tableau calculus for L_T put to a *semantic* (and not proof theoretic) use. The most important difference with respect to Smullyan's signed tableau rules for first order logic ([50]) is the addition of rules for the truth predicate. The assertoric rules for the truth predicate, for conjunction and for negation were already displayed above.

In a *closure game*, there are two players, called $\sqcup$ and $\sqcap$. Player $\sqcup$ controls all AD sentences of *disjunctive type* and player $\sqcap$ controls all sentences of *conjunctive type*. Sentences of form $A_{\alpha\vee\beta}, D_{\alpha\wedge\beta}, A_{\exists\phi(x)}, D_{\forall\phi(x)}$ are of disjunctive type, all others of conjunctive type. A *strategy* of a player is a mapping of each AD sentence X_σ that is in his control to exactly one of the *immediate successors* of X_σ, as specified by the assertoric rule applicable to X_σ. A few examples suffice to illustrate the notion of a strategy. The immediate successors of $A_{\alpha\wedge\beta}$ are A_α and A_β and, as $A_{\alpha\wedge\beta}$ is of conjunctive type, a strategy of player $\sqcap$ maps $A_{\alpha\wedge\beta}$ to either A_α or A_β. As $A_{T(\overline{\sigma})}$ has only one immediate successor, A_σ, every strategy of player $\sqcap$ must map $A_{T(\overline{\sigma})}$ to A_σ. A strategy for player $\sqcup$, who controls $D_{\alpha\wedge\beta}$, maps $D_{\alpha\wedge\beta}$ to either D_α or D_β.

With f a strategy for player $\sqcup$, g a strategy for player $\sqcap$ and with X_σ an arbitrary AD sentence, the tuple (X_σ, f, g) defines an *expansion of* X_σ. In general, an expansion of X_σ is an infinite[20] sequence of AD sentences whose first element is X_σ and whose successor relation respects the assertoric rules. As an example, here is the expansion of $A_{\neg T(\lambda)}$, i.e., of an assertion of the Liar:

$$A_{\neg T(\lambda)}, D_{T(\lambda)}, D_{\neg T(\lambda)}, A_{T(\lambda)}, A_{\neg T(\lambda)} \dots \tag{6.5}$$

Indeed, $A_{\neg T(\lambda)}$ has only one expansion and so, in *the closure game for* $A_{\neg T(\lambda)}$, none of the players can influence the expansion of $A_{\neg T(\lambda)}$ that is realized. In general, an AD sentence X_σ may have (infinitely) many expansions, each of which is realized by some strategy pair (f, g) of our players. For instance, $A_{P(c_1)\wedge P(c_2)}$, where $P(c_1)$ and $P(c_2)$ are atomic sentences of L, has two expansions and, in *the closure game for* $A_{P(c_1)\wedge P(c_2)}$, player $\sqcap$ can determine which one is realized. By setting $g(A_{P(c_1)\wedge P(c_2)}) = A_{P(c_1)}$, player $\sqcap$ ensures that expansion (6.6) is realized, while $g(A_{P(c_1)\wedge P(c_2)}) = A_{P(c_2)}$ realizes expansion (6.7).

[20] Whenever an expansion "hits" a signed atomic sentence of L it keeps on repeating it indefinitely.

$$A_{P(c_1)\wedge P(c_2)}, A_{P(c_1)}, A_{P(c_1)}, A_{P(c_1)}, \ldots \tag{6.6}$$

$$A_{P(c_1)\wedge P(c_2)}, A_{P(c_2)}, A_{P(c_2)}, A_{P(c_2)}, \ldots \tag{6.7}$$

We will write $\mathsf{exp}(X_\sigma, f, g)$ to denote the expansion of X_σ that is induced by strategies f (for player $\sqcup$) and g (for player $\sqcap$).

With M a ground model, a *closure condition* $\dagger(M)$ is a bipartition $\{O^\dagger_M, C^\dagger_M\}$ of the set of all expansions in M. The sets $O^\dagger_M$ and $C^\dagger_M$ consists of all open and all closed expansions in M respectively. In a closure game for X_σ played under closure conditions $\dagger(M)$, player $\sqcup$ tries to pick his strategy f in such a way that the expansion of X_σ that is realized will be contained in $O^\dagger_M$. We will write $O^\dagger_M(X_\sigma)$, and say that X_σ *is open relative to* $\dagger(M)$, to indicate that player $\sqcup$ has a strategy which *ensures* that the expansion of X_σ ends up in $O^\dagger_M$. That is:

$$O^\dagger_M(X_\sigma) \Leftrightarrow \exists f \forall g \;\; \mathsf{exp}(X_\sigma, f, g) \in O^\dagger_M \tag{6.8}$$

X_σ is *closed relative to* $\dagger(M)$, denoted $C^\dagger_M(X_\sigma)$, just in case not $O^\dagger_M(X_\sigma)$. As specified by (6.8), a closure condition for expansions induces a closure condition for AD sentences. The closure condition for AD sentences is used to induce a valuation for L_T, denoted $\mathcal{V}^\dagger_M$:

$$\mathcal{V}^\dagger_M(\sigma) = \begin{cases} \mathbf{a} := (1,0), & O^\dagger_M(A_\sigma) \text{ and } C^\dagger_M(D_\sigma); \\ \mathbf{b} := (1,1), & O^\dagger_M(A_\sigma) \text{ and } O^\dagger_M(D_\sigma); \\ \mathbf{n} := (0,0), & C^\dagger_M(A_\sigma) \text{ and } C^\dagger_M(D_\sigma); \\ \mathbf{d} := (0,1), & C^\dagger_M(A_\sigma) \text{ and } O^\dagger_M(D_\sigma). \end{cases} \tag{6.9}$$

In general $\mathcal{V}^\dagger_M$ may, but need not have, a range of four values. The intuitive interpretation of the functions that are induced by the method of closure games is an assertoric one. For instance, $\mathcal{V}^\dagger_M(\sigma) = \mathbf{a}$ indicates that it is allowed to assert, but not to deny, sentence σ in ground model M according to the norms for assertion and denial that are specified by $\dagger(M)$.

The method of closure games allows us to characterize all 3- and 4-valued Strong Kleene fixed point valuations in a uniform manner as shown in [63] Here is a rough sketch of the characterization. For each expansion exp, its *successor expansion* exp', is obtained by deleting the first term of exp. For instance, (6.10) is the successor expansion of (6.6).

$$A_{P(c_1)}, A_{P(c_1)}, A_{P(c_1)}, \ldots \tag{6.10}$$

A closure condition $\dagger(M) = \{O^\dagger_M, C^\dagger_M\}$ satisfies the *Stable Judgement Constraint* (SJC), just in case for every expansion exp we have that:

$$\mathsf{SJC}: \qquad \mathsf{exp} \in C^\dagger_M \Leftrightarrow \mathsf{exp}' \in C^\dagger_M$$

If a closure condition $\dagger(M)$ satisfies SJC, the judgement of $\dagger$ as to whether an expansion is open or closed is stable, in the sense that it does not change along the expansion. It can be shown (see XXX) that whenever a closure condition satisfies SJC, it induces a 2-, 3- or 4-valued Strong Kleene theory of truth and, conversely, if V_M is a 2-, 3- or 4-valued theory of truth, it can be induced via

the method of closure games by closure conditions that satisfy SJC[21].

Inducing $\mathcal{K}_M$ via the Method of Closure games. We say that an expansion is *grounded* just in case it hits a signed atomic sentence of L and *ungrounded* otherwise. Thus, expansions (6.6) and (6.7) are grounded, whereas (6.5) is ungrounded. We say that an expansion exp is *grounded and correct in* M just in case exp is grounded and, with X_σ the (unique) signed atomic sentence of L that occurs on exp, we have that:

- $(X_\sigma = A_\sigma$ and $\mathcal{C}_M(\sigma) = \mathbf{a})$ or $(X_\sigma = D_\sigma$ and $\mathcal{C}_M(\sigma) = \mathbf{d})$.

As shown in [63], we can induce $\mathcal{K}_M$ as follows.

Theorem 6.3 Inducing $\mathcal{K}_M$ Consider the following gr(oundedness) closure conditions $\{O^{gr}_M, C^{gr}_M\}$, where:

$$\mathsf{exp} \in C^{gr}_M \Leftrightarrow \mathsf{exp} \text{ is ungrounded or grounded and incorrect in } M$$

Equivalently, we have that:

$$\mathsf{exp} \in O^{gr}_M \Leftrightarrow \mathsf{exp} \text{ is grounded and correct in } M$$

The function $\mathcal{V}^{gr}_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$ which is induced by the groundedness closure conditions is equivalent—modulo a translation of $\mathbf{u}$ as $\mathbf{n}$—to $\mathcal{K}_M$.
Proof: See [63].

The Method of Closure Games and (N)OA models.
By a *NOA model*, we mean a triple $\mathcal{M} = (M, \mathbf{Kn}^+, \mathbf{Kn}^-)$, where:

1. M is a ground model, inducing the classical $\mathcal{C}_M : Sen(L) \to \{0, 1\}$.
2. $\mathbf{Kn}^+ \subseteq Sen(L)$
3. $\sigma \in \mathbf{Kn}^+ \Rightarrow \mathcal{C}_M(\sigma) = 1$
4. $\mathbf{Kn}^+$ is closed under (classical) logical consequence.
5. $\mathbf{Kn}^- = \{\sigma \mid \neg\sigma \in \mathbf{Kn}^+\}$

Thus, $\mathbf{Kn}^+$ are the sentences of L (condition 2) that are known by a *logically* omniscient agent (condition 4), where knowledge is factive (condition 3). $\mathbf{Kn}^-$ is the set of sentences of L whose negation is known (condition 5). An *OA model* is a *NOA* model $\mathcal{M}$ in which $\mathbf{Kn}^+ \cup \mathbf{Kn}^- = Sen(L)$. Indeed, for each ground model M, there is exactly one *OA* model $\mathcal{M} = (M, \mathbf{Kn}^+, \mathbf{Kn}^-)$.

With $\mathcal{M} = (M, \mathbf{Kn}^+, \mathbf{Kn}^-)$ a *NOA* model, we may say that an expansion exp is *known to be correct*, just in case there occurs an element X_σ on exp such that:

[21] By a 2-valued Strong Kleene theory of truth, we mean a classical valuation which respects the identity of truth and the ground model M. If M contains Liar sentences (or their ilk), there are no 2-valued Strong Kleene theories of truth over M. Further, when $\dagger(M)$ satisfies SJC it is not guaranteed that the valuation induced by $\dagger(M)$ respects the ground model M. However, a further (obvious) constraint on closure conditions (called the *world respecting constraint* in [63]) can be added which ensures that the induced valuation does respect M.

- $(X_\sigma = A_\sigma$ and $\sigma \in \mathbf{Kn}^+)$ or $(X_\sigma = D_\sigma$ and $\sigma \in \mathbf{Kn}^-)$.

Consider the following *kn(owledge)* closure conditions $\{O^{kn}_{\mathcal{M}}, C^{kn}_{\mathcal{M}}\}$ for a NOA model $\mathcal{M}$:

$$\mathsf{exp} \in O^{kn}_{\mathcal{M}} \Leftrightarrow \mathsf{exp} \text{ is known to be correct}$$

By playing a closure game under the knowledge closure conditions in a NOA model $\mathcal{M}$ we induce the valuation function $\mathcal{V}^{kn}_{\mathcal{M}}$ which, *when* $\mathcal{M} = (M, \mathbf{Kn}^+, \mathbf{Kn}^-)$ *is an OA model*, is identical to $\mathcal{K}_M$, as a little reflection on the groundedness and knowledge closure conditions shows.

Now let $\mathcal{M} = (M, \mathbf{Kn}^+, \mathbf{Kn}^-)$ be a (genuine) NOA model, i.e., there are sentences of L which are neither contained in $\mathbf{Kn}^+$ nor in $\mathbf{Kn}^-$. For sake of definiteness, suppose that $H(c)$ is such a sentence, where c denotes Julias Ceasar and where $H(\cdot)$ expresses the property of having exactly 12 hairs on one's big toe when crossing the Rubicon, i.e., $H(c)$ represents sentence (6.4). Consider what happens if we play a closure game under the knowledge closure conditions in this NOA model $\mathcal{M}$. As $H(c)$ is neither contained in $\mathbf{Kn}^+$ nor in $\mathbf{Kn}^-$, we get that $\mathcal{V}^{kn}_{\mathcal{M}}(H(c)) = \mathbf{n}$, which is "as it should be". Further, as $\mathbf{Kn}^+ \subseteq Sen(L)$ is closed under classical logical consequence, it follows that $H(c) \vee \neg H(c) \in \mathbf{Kn}^+$, from which it follows that $\mathcal{V}^{kn}_{\mathcal{M}}(H(c) \vee \neg H(c)) = \mathbf{a}$. As we are modeling the knowledge of a logical omniscient agent, this is also "as it should be". Although $H(c)$ can neither be asserted nor denied, we surely want its T-sentence to be assertible. That is, we want to have that $\mathcal{V}^{kn}_{\mathcal{M}}(T([H(c)]) \leftrightarrow H(c)) = \mathbf{a}$. However, on the present account, we have that $\mathcal{V}^{kn}_{\mathcal{M}}(T([H(c)]) \leftrightarrow H(c)) = \mathbf{n}$. To see why, it is convenient to consider the (derived) assertoric rules for material implication:

$\rightarrow$	$\dfrac{A_{(\alpha\rightarrow\beta)}}{D_\alpha \mid A_\beta}$	$\dfrac{D_{(\alpha\rightarrow\beta)}}{A_\alpha, D_\beta}$

We will show that $\mathcal{V}^{kn}_{\mathcal{M}}(T([H(c)]) \rightarrow H(c)) = \mathbf{n}$, a similar argument reveals that $\mathcal{V}^{kn}_{M}(T([H(c)]) \leftrightarrow H(c)) = \mathbf{n}$. In the closure game for $A_{T([H(c)])\rightarrow H(c)}$, player $\sqcup$ chooses whether he picks his strategy f such that $f(A_{T([H(c)])\rightarrow H(c)}) = D_{T([H(c)])}$ or $f(A_{T([H(c)])\rightarrow H(c)}) = A_{H(c)}$. His choices induce, respectively, the following expansions of $A_{T([H(c)])\rightarrow H(c)}$.

$$A_{T([H(c)])\rightarrow H(c)}, D_{T([H(c)])}, D_{H(c)}, D_{H(c)}, \ldots \tag{6.11}$$

$$A_{T([H(c)])\rightarrow H(c)}, A_{H(c)}, A_{H(c)}, A_{H(c)}, A_{H(c)} \ldots \tag{6.12}$$

Both expansions are closed according to the knowledge closure conditions, and so we have that $C^{kn}_{\mathcal{M}}(A_{T([H(c)])\rightarrow H(c)})$. A similar argument reveals that $C^{kn}_{\mathcal{M}}(D_{T([H(c)])\rightarrow H(c)})$ and so we have that $\mathcal{V}^{kn}_{\mathcal{M}}(T([H(c)]) \rightarrow H(c)) = \mathbf{n}$. The undesirable valuation of grounded T-sentences such as those of $H(c)$ is due to the fact that $\mathbf{Kn}^+$ is logically closed under the rules of classical logic, which do not take into account the inferential rules for truth.

However, the undesirable valuation of grounded T-sentences is easily repaired. We can do so by closing off $\mathbf{Kn}^+$ under the inferential truth rules or, equivalently, by posing further conditions under which an AD sentence X_σ is open in a closure game. Here, we opt for the latter. We may say that an AD sentence X_σ is *logically open*, just in case, for some $\alpha \in L$, the following two conditions hold:

1. $\exists f \forall g$ $\mathsf{exp}(X_\sigma, f, g)$ contains A_α

2. $\exists f \forall g$ $\mathsf{exp}(X_\sigma, f, g)$ contains D_α

Using the notion of logical openess, we may define the adjusted knowledge closure conditions for AD sentence as follows. An AD sentence X_σ is said to be *open according to the adjusted knowledge closure conditions in NOA model* $\mathcal{M}$, denoted $O^{kn\star}_{\mathcal{M}}(X_\sigma)$, just in case:

$$O^{kn}_{\mathcal{M}}(X_\sigma) \text{ or } X_\sigma \text{ is logically open.}$$

When we have that not $O^{kn\star}_{\mathcal{M}}(X_\sigma)$, we say that X_σ is closed according to the adjusted knowledge closure conditions, denoted $C^{kn\star}_{\mathcal{M}}(X_\sigma)$. As an example, we have that $O^{kn\star}_{\mathcal{M}}(A_{T([H(c)])\rightarrow H(c)})$. For, although we have that $C^{kn}_{\mathcal{M}}(A_{T([H(c)])\rightarrow H(c)})$, $A_{T([H(c)])\rightarrow H(c)}$ is logically open.

For any NOA model $\mathcal{M}$, we let $\mathcal{V}^{kn\star}_{\mathcal{M}} : Sen(L_T) \rightarrow \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$ be the valuation function that is induced by $O^{kn\star}$ and $C^{kn\star}$ in accordance with the schema of (6.9). We have that:

Proposition 6.1 On the valuation function $\mathcal{V}^{kn\star}_{\mathcal{M}}$
Let $\mathcal{M} = (M, \mathbf{Kn}^+, \mathbf{Kn}^-)$ be a NOA model and let $\mathcal{V}^{kn\star}_{\mathcal{M}} : Sen(L_T) \rightarrow \{\mathbf{a}, \mathbf{n}, \mathbf{d}\}$ be the associated valuation function. We have that:

1. $\{\mathbf{a}, \mathbf{d}\}$ is a strong assertoric pair of $\mathcal{V}^{kn\star}_{\mathcal{M}}$.

2. $\mathcal{K}_M(\sigma) \neq \mathbf{u} \Rightarrow \mathcal{V}^{kn\star}_{\mathcal{M}}(T(\overline{\sigma}) \leftrightarrow \sigma) = \mathbf{a}$.

3. $\mathcal{K}_M(\sigma) = \mathbf{u} \Rightarrow \mathcal{V}^{kn\star}_{\mathcal{M}}(\sigma) = \mathbf{n}$.

Proof: By an inspection of the definitions. □

Clearly, $\mathcal{V}^{kn\star}_{\mathcal{M}}$ is not a compositional valuation function. For instance, with $H(c)$ as above, we have that $\mathcal{V}^{kn\star}_{\mathcal{M}}(H(c)) = \mathcal{V}^{kn\star}_{\mathcal{M}}(\neg H(c)) = \mathbf{n}$, whereas $\mathcal{V}^{kn\star}_{\mathcal{M}}(H(c) \vee \neg H(c)) = \mathbf{a}$. This is not something to be excused for, but rather a straightforward consequence of the fact that we are modeling the assertoric possibilities of a non omniscient agent under the knowledge norm for assertion. The third clause of Proposition 6.1 tells us that according to $\mathcal{V}^{kn\star}_{\mathcal{M}}$, ungrounded sentences are neither assertible nor deniable. Hence, $\mathcal{V}^{kn\star}_{\mathcal{M}}$ shares its judgements with respect to ungrounded sentence with $\mathcal{K}_M$ and so, as before, these judgements can be "overruled" via $\mathcal{K}^+_M$. Doing so, we obtain the theory $\mathcal{V}^{\star}_{\mathcal{M}}$:

$$\mathcal{V}^{\star}_{\mathcal{M}}(\sigma) = \begin{cases} \mathbf{a}_g, & \mathcal{V}^{kn\star}_{\mathcal{M}}(\sigma) = \mathbf{a} \\ \mathbf{a}_i, & \mathcal{K}_M(\sigma) = \mathbf{u}, \mathcal{K}^+_M(\sigma) = \mathbf{a} \\ \mathbf{n}_g, & \mathcal{K}_M(\sigma) \neq \mathbf{u}, \mathcal{V}^{kn\star}_{\mathcal{M}}(\sigma) = \mathbf{n} \\ \mathbf{e}, & \mathcal{K}^+_M(\sigma) = \mathbf{e} \\ \mathbf{d}_i, & \mathcal{K}_M(\sigma) = \mathbf{u}, \mathcal{K}^+_M(\sigma) = \mathbf{d} \\ \mathbf{d}_g, & \mathcal{V}^{kn\star}_{\mathcal{M}}(\sigma) = \mathbf{d} \end{cases}$$

Just as $\mathcal{K}^{\mathbf{5}}_M$, $\mathcal{V}^{\star}_{\mathcal{M}}$ satisfies **AD** and violates **MGBD**. However, a little care has to be taken when confronting $\mathcal{V}^{\star}_{\mathcal{M}}$ with those desiderata, as $\mathcal{V}^{\star}_{\mathcal{M}}$ is defined relative to a NOA model $\mathcal{M}$ and the desiderata are formulated in terms of ground models M. It doesn't make sense to define the notion of no-vicious reference relative to a NOA model. For, suppose that one has a NOA model

in which $H(c)$ is a before and in which there is a non-quotational constant d such that $I(d) = H(c)$. As $\mathcal{V}^{\star}_{\mathcal{M}}(H(c)) = \mathbf{n}_g$, there is reference to a "non-classical" sentence in this NOA model and so there would be vicious reference. But clearly, that's not the right verdict, as the non-classicality of $H(c)$ does not arise from the semantics of the truth predicate. Similarly, it doesn't make sense to define the notion of truth behaving as a classical concept relative to a NOA model. For, in a NOA model where $H(c)$ is as before, truth would not behave like a classical concept according to such a definition. But again, this is due to the "non-classicality" of $H(c)$, which is not explained by the semantics of the truth predicate[22]. In contrast, it does make sense to define the strong assertibility of all T-sentences relative to a NOA model. Although a lack of knowledge may refrain us from asserting or denying $H(c)$, this should not refrain us from asserting its T-sentence. Indeed, this feature *is* to be explained via the semantics of the truth predicate.

In other words, with respect to any NOA model $\mathcal{M} = (M, \mathbf{Kn}^+, \mathbf{Kn}^-)$, the notion of no vicious reference according to $\mathcal{V}^{\star}_{\mathcal{M}}$ and the notion of truth behaving classical according to $\mathcal{V}^{\star}_{\mathcal{M}}$ are to be defined[23] via $\mathcal{K}^{\mathbf{5}}_{M}$. The notion of the strong assertibility of all T-sentences, however, can be defined directly in terms of $\mathcal{V}^{\star}_{\mathcal{M}}$. Doing so, we see that, in light of Proposition 6.1, $\mathcal{V}^{\star}_{\mathcal{M}}$ satisfies **AD** and violates **MGBD**. Hence, the main point of this paper is not threatened by **Obj**, even when we accept its appeal to the knowledge account of assertion.

6.7 Further remarks on desiderata for theories of truth

6.7.1 A third desideratum

We remarked that $\mathcal{K}$ valuates the law of excluded middle, LEM, as *ungrounded* in a clean ground model. The fact that it does so has a clear intuitive explanation. LEM is a sentence of form $\forall x \phi(x)$ and the quantifiers of L_T range over all sentences of L_T. Thus, LEM quantifies over itself and, intuitively, says (amongst others) of itself that it obeys the law of excluded middle. In more detail, an assertion of LEM commits one (amongst others) to the assertion of $T([\mathsf{LEM}]) \vee \neg T([\mathsf{LEM}])$. To discharge a commitment to a disjunction, we have[24] to assert one of the disjuncts, and it is clear that it is $T(\mathsf{LEM})$ which should be asserted in this case. But an assertion of $T(\mathsf{LEM})$ comes down to an assertion of LEM itself. Thus, an assertion of LEM exhibits, intuitively, a kind of circularity. Observe that the (intuitive) argument for the ungroundedness of LEM just given holds in any ground model, and not just in clean ones. That is: $\mathcal{K}_M(\mathsf{LEM}) = \mathbf{u}$ for *any* ground model M. We say that according to $\mathcal{K}$, LEM is *essentially* ungrounded. Thus, we have that:

Proposition 6.2 For *every* **ground model M: $\mathcal{K}$ and $\mathcal{K}^{\mathbf{5}}$ dictate that truth does** *not* **behave like a classical concept in M.**

[22] Compare the remarks on category mistakes in Section 4.1. On such a view, the non-classicality of 'snow is true' *is* explained by the semantics of the truth predicate: the fact that the truth predicate is only properly applied to sentences explains that to attribute truth to snow is to make a mistake.

[23] Or, equivalently, by the function $\mathcal{V}^{\star}_{\mathcal{M}'}$ where $\mathcal{M}'$ is the unique OA model over M.

[24] As we assume a Strong Kleene interpretation of the logical connectives.

Proof: From the observation that $\mathcal{K}$ valuates the law of excluded middle as **u** in every ground model. □

In light of Proposition 6.2, which quantifies over all ground models, we may say that according to $\mathcal{K}$ and $\mathcal{K}^{\mathbf{5}}$ truth is *essentially* a non-classical concept[25]. I do not see anything bad in a theory of truth according to which truth is essentially non-classical, as the source of this non-classicality can be explained, in a intuitively appealing manner, by the essential ungroundedness of LEM. In fact, a proponent of $\mathcal{K}^{\mathbf{5}}$ may argue that the implicit circularity involved in the assertion of LEM as described above, testifies that truth *is* essentially a non-classical concept. According to $\mathcal{K}^{\mathbf{5}}$, LEM is (strongly) assertible in a clean ground model for non-classical reasons. For someone who thinks that truth is essentially a non-classical concept, it seems reasonable to impose the following **N**on **C**lassicality **D**esideratum on theories of truth:

NCD In any ground model M, a theory of truth **T** should dictate that truth does not behave like a classical concept in M.

Indeed, any theory which satisfies **MGBD** violates **NCD**. When both **AD** and **NCD** are accepted as desiderata for theories of truth, $\mathcal{K}^{\mathbf{5}}$ does better than *all* theories of Theorem 6.1. In particular, it does better than its constituent theories $\mathcal{K}$, which violates **AD**, and $\mathcal{K}^{+}$, which violates **NCD**. In the next subsection, we will see another example of a theory which satisfies both **AD** and **NCD**.

It is instructive to compare the considered reaction to the behavior of LEM in clean ground models to a reaction[26] that is considered by [32].

> The story about grounding might trump any intuitions that blame truth's nonclassical behaviour on vicious reference. Indeed, we could go further and insist that there actually is vicious reference in this simple ground model after all, since the quote name $\forall x T(x) \vee \neg T(x)$ viciously refers to the ungrounded sentence $\forall x T(x) \vee \neg T(x)$. [...] But it is a kind of vicious reference that has no apparent relationship to the kind of vicious reference that has traditionally been seen as a source of paradox or pathology. (Kremer [32, p362])

Thus, Kremer suggests that a proponent of $\mathcal{K}$—in light of $\mathcal{K}$'s failure of **MGBD**—may consider the notion of groundedness so important that he is willing to admit that even in clean ground models there is vicious reference. This is a different reaction to (amongst others) LEM's behavior than the one that is put forward by a **NCD** proponent. For, a **NCD** proponent is happy to grant that there is no vicious reference in clean ground models. Still, the absence of vicious reference need not force truth to behave classically: for a **NCD** proponent, truth is *essentially* a non-classical notion.

[25] For some theories **T**, the violation of **MGBD** is dependent on ground model under consideration. For instance, $\mathcal{S}$, the minimal fixed point theory based on the Supervaluation theory satisfies **MGBD** with respect to a clean ground model, as the reader may wish to verify. $\mathcal{S}$ violates **MGBD** though, as shown in example 5.10 of [31].

[26] Kremer also considers, on page 362, a reaction that is in line with ours: 'So, despite the apparent absence of vicious reference, LEM seems ungrounded in our intuitive sense.'

6.7.2 The theory $\mathbb{V}^{8+}$

There are more theories than $\mathcal{K}^5$ which satisfy both **AD** and **NCD**. An example is the eight valued generalized Strong Kleene theory $\mathbb{V}^{8+}$, that was defined in [63]. The semantics of $\mathbb{V}^{8+}$ is described via a generalization of the Strong Kleene semantics for four valued theories of truth. The only distinction between the semantics of a four valued Strong Kleene theory and the semantics of $\mathbb{V}^{8+}$ is due to the fact that negation acts as "a swap operation" on three pairs of semantic values. Besides that distinction, the semantics is Strong Kleene, and can be described in terms of the lattice $\mathbf{8}^+_{\leq}$, whose Hasse diagram is depicted below.

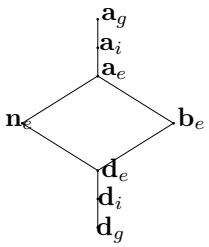


Figure 6.1: Hasse diagram of $\mathbf{8}^+_{\leq}$, the lattice of $\mathbb{V}^{8+}$.

Conjunction and disjunction act as meet and join in $\mathbf{8}^+_{\leq}$, and universal and existential quantification act as generalized conjunction and disjunction. Negation acts as the identity operation on $\mathbf{b}_e$ and $\mathbf{n}_e$, but also, it interchanges $\mathbf{a}_x$ for $\mathbf{d}_x$, where $x \in \{g, i, e\}$ indicates the assertoric sense under consideration: *g*rounded, *i*ntrinsic or *e*xtrinsic. $\mathbb{V}^{8+}$ can be defined in terms of $\mathcal{K}^5$ and a four valued theory $\mathbb{V}^{4+}$ that is defined in [63] using the method of closure games. The details of the definition do not matter for our purposes. For any ground model M, we have that $\mathbb{V}^{4+}_M : Sen(L_T) \to \{\mathbf{a}, \mathbf{b}, \mathbf{n}, \mathbf{d}\}$, where $\mathbf{b}$ and $\mathbf{n}$ are interpreted as *both assertible and deniable*, and *neither assertible nor deniable* respectively. The Truthteller is an example of a sentence which is valuated as $\mathbf{b}$, while the Liar is an example of a sentence which is valuated as $\mathbf{n}$. The relation between $\mathcal{K}^5$ and $\mathbb{V}^{8+}$ is as follows, where $\mathbf{x} \in \{\mathbf{a}, \mathbf{d}\}$:

$$\mathcal{K}^5_M(\sigma) = \mathbf{x}_g \Leftrightarrow \mathbb{V}^{8+}_M(\sigma) = \mathbf{x}_g, \qquad \mathcal{K}^5_M(\sigma) = \mathbf{x}_i \Leftrightarrow \mathbb{V}^{8+}_M(\sigma) = \mathbf{x}_i \tag{6.13}$$

From equation (6.13) it follows that $\mathbb{V}^{8+}$ violates **MGBD** and that it satisfies both **AD** and **NCD**. $\mathbb{V}^{8+}$ uses $\mathbb{V}^{4+}$ to impose a fine grained assertoric structure on the sentences that are valuated as $\mathbf{e}$ by $\mathcal{K}^5$. That such is possible in an (eight valued) generalized compositional way is a consequence of the fact that:

$$\mathcal{K}_M(\sigma) = \mathbf{a} \Rightarrow \mathcal{K}^+_M(\sigma) = \mathbf{a} \Rightarrow \mathbb{V}^{4+}_M(\sigma) = \mathbf{a} \tag{6.14}$$

From the perspective of $\mathbb{V}^{8+}$, $\mathcal{K}$, $\mathcal{K}^+$ and $\mathbb{V}^{4+}$ model three distinct assertoric senses. Here is a table that gives some intuitions concerning $\mathbb{V}^{8+}$.

σ	$\mathcal{K}_M(\sigma)$	$\mathcal{K}^+_M(\sigma)$	$\mathbb{V}^{\mathbf{4}+}_M(\sigma)$	$\mathbb{V}^{\mathbf{8}+}_M(\sigma)$
$c \approx c$	$\mathbf{a}$	$\mathbf{a}$	$\mathbf{a}$	$\mathbf{a}_g$
$T(\eta) \vee \neg T(\eta)$	$\mathbf{u}$	$\mathbf{a}$	$\mathbf{a}$	$\mathbf{a}_i$
$\neg T(\lambda) \vee T(\tau)$	$\mathbf{u}$	$\mathbf{e}$	$\mathbf{a}$	$\mathbf{a}_e$
$T(\tau)$	$\mathbf{u}$	$\mathbf{e}$	$\mathbf{b}$	$\mathbf{b}_e$
$\neg T(\lambda)$	$\mathbf{u}$	$\mathbf{e}$	$\mathbf{n}$	$\mathbf{n}_e$
$\neg T(\lambda) \wedge T(\tau)$	$\mathbf{u}$	$\mathbf{e}$	$\mathbf{d}$	$\mathbf{d}_e$
$T(\theta) \wedge \neg T(\theta)$	$\mathbf{u}$	$\mathbf{e}$	$\mathbf{a}$	$\mathbf{d}_i$
$c \not\approx c$	$\mathbf{d}$	$\mathbf{d}$	$\mathbf{d}$	$\mathbf{d}_g$

$\mathbb{V}^{\mathbf{8}+}$ is another example of a theory of truth which satisfies both **AD** and **NCD**.

6.7.3 To sum up

We proposed a formal and theory-relative desideratum for theories of truth, **AD**, which is, from a formal point of view, closely related to Kremer's **MGBD**. From a philosophical point of view, however, **AD** is fundamentally distinct from **MGBD**. We argued that **AD** is preferable over **MGBD**. We gave examples of theories of truth, $\mathcal{K}^{\mathbf{5}}$ and $\mathbb{V}^{\mathbf{8}+}$, which violate **MGBD** and satisfy **AD**. Also, we saw that the law of excluded middle suggest that truth may be *essentially* a non-classical notion, which led to the formulation of **NCD**, a third desideratum for theories of truth. Both $\mathcal{K}^{\mathbf{5}}$ and $\mathbb{V}^{\mathbf{8}+}$ satisfy **NCD** and, interestingly, any theory of truth which satisfies **MGBD** violates **NCD**.

Chapter 7

Strict-Tolerant Tableaux for Strong Kleene Truth

7.1 Abstract

We discuss four distinct semantic consequence relations which are based on Strong Kleene theories of truth and which generalize the notion of classical consequence to 3-valued logics. Then we set up a uniform signed tableau calculus (the *strict-tolerant calculus*) which we show to be sound and complete with respect to each of the four semantic consequence relations. The signs employed by our calculus are A^s, D^s, A^t and D^t which indicate a *strict assertion*, *strict denial*, *tolerant assertion* and *tolerant denial* respectively. Recently, Ripley applied the strict-tolerant account of assertion and denial (originally developed by Cobreros et all. to bear on vagueness) to develop a new approach to truth and alethic paradox, which we call the *Strict Tolerant Conception of Truth* (STCT). The paper aims to contribute to our understanding of STCT in at least three ways. First, by developing the strict-tolerant calculus. Second, by developing a semantic version of the strict-tolerant calculus (*assertoric semantics*) which informs us about the (strict-tolerant) assertoric possibilities relative to a fixed *ground model*. Third, by showing that the strict-tolerant calculus and assertoric semantics jointly suggest that STCT's claim that the strict-tolerant distinction is not a primitive one, as"the strict and tolerant can be understood in terms of one another", has to be reconsidered. The paper concludes with a methodological comparison between the strict-tolerant calculus and other calculi that are also sound and complete with respect to (some of the) semantic consequence relations based on Strong Kleene theories of truth.

7.2 Introduction

Classical logic recognizes two semantic values, which we call 1, the *designated value* and 0, the *anti-designated* value. Classical consequence can be defined in terms of *preservation of designated value*. Working in a multiple-conclusion setting, a premise set Γ is said to (classically) *entail* a conclusion set Δ, denoted

$\Gamma \models^{cl} \Delta$, just in case, in passing from Γ to Δ, the value 1 is preserved, i.e.:

$$\text{All } \alpha \in \Gamma \text{ are valuated as } 1 \Rightarrow \text{some } \beta \in \Delta \text{ is valuated as } 1. \tag{7.1}$$

However, we may also characterize the classical consequence relation in terms of the preservation of "non anti-designated value". For, $\Gamma \models^{cl} \Delta$ just in case:

$$\text{All } \alpha \in \Gamma \text{ are valuated as non-0} \Rightarrow \text{some } \beta \in \Delta \text{ is valuated as non-0.} \tag{7.2}$$

Here are two other characterizations of classical consequence, both equivalent to (7.1) and (7.2). We have that $\Gamma \models^{cl} \Delta$ just in case:

$$\text{All } \alpha \in \Gamma \text{ are valuated as } 1 \Rightarrow \text{some } \beta \in \Delta \text{ is valuated as non-0.} \tag{7.3}$$

$$\text{All } \alpha \in \Gamma \text{ are valuated as non-0} \Rightarrow \text{some } \beta \in \Delta \text{ is valuated as } 1. \tag{7.4}$$

Although (7.1), (7.2), (7.3) and (7.4) are all equivalent in classical logic, they come apart when we move to a non-classical setting in which there are more than two semantic values. In this paper, we will be concerned with a particular such non-classical setting; we will be concerned with consequence relations that are induced by *Strong Kleene fixed point valuations* of a language L_T, containing a distinguished truth predicate symbol T. Such consequence relations we call *fixed point consequence* relations.

Below, we will define four (primitive) fixed point consequence relations for L_T that correspond to formulas (7.1), (7.2), (7.3) and (7.4) respectively. We will do so in terms of *strict* and *tolerant* assertion and denial, a distinction originally due to Cobreros et al [11]. In section 7.3, we define a uniform tableau calculus, called the *strict-tolerant calculus* that can be used to define syntactic consequence relations that are sound and complete with respect to these four fixed point consequence relations. In section 7.4, we develop a semantic version of the strict-tolerant calculus, called *assertoric semantics*, that can be used to determine the (strict-tolerant) assertoric status of L_T sentences relative to a fixed *ground model*, i.e., relative to a fixed interpretation of the truth-free fragment of L_T. In Section 7.5, we discuss a recent philosophical approach to truth, due to Ripley [46], which we call the *Strict Tolerant Conception of Truth* (STCT). As STCT heavily relies on the strict-tolerant distinction, it is interesting to see how the technical of Section 7.3 and 7.4 relate to STCT. We will argue that the strict-tolerant calculus and assertoric semantics jointly suggest that, pace Ripley, the strict-tolerant distinction is a primitive one. In other words, we argue that the strict and tolerant cannot be understood in terms of one another. Section 7.6 highlights the most important distinctions between the strict-tolerant calculus and other calculi that are also sound and complete with respect to (some of the) fixed point consequence relations. Section 7.6 concludes. An appendix gives a proof of a theorem of Section 7.4.

The remainder of this introduction defines and discusses the fixed point consequence relations in strict-tolerant terms (Section 7.2.1) and explains, in sections 7.2.2, 7.2.3 and 7.3.4, the essential ideas of sections 7.3, 7.4 and 7.5 respectively.

7.2.1 Fixed point consequence in strict-tolerant terms

Before we introduce the strict-tolerant slang and the four fixed point consequence relations in terms of it, we define the notion of a (Strong Kleene, but let

that be understood) *fixed point valuation of* L_T. L_T is a first order language without function symbols, with identity ($\approx$), a truth predicate (T) and a quotational constant symbol ($[\sigma]$) for each sentence σ of L_T. As there are no function symbols around, the set of closed terms of L_T is given by $Con(L_T)$, i.e., the set of all (quotational and non-quotational) constant symbols. A *ground model* $M = (D, I)$ is a classical interpretation for L, the truth-free fragment of L_T, (in which $\approx$ is interpreted as identity and) such that:

$$Sen(L_T) \subseteq D, \qquad I([\sigma]) = \sigma$$

With $M = (D, I)$ a ground model and $\sigma \in Sen(L_T)$ we will use $\overline{\sigma}$ to denote an arbitrary constant (quotational or not) of L_T such that $I(\overline{\sigma}) = \sigma$. A ground model M equips L with a classical valuation $\mathcal{C}_M : Sen(L) \to \{0, 1\}$. A *fixed point valuation of* L_T relative to ground model M is a function $V_M : Sen(L_T) \to \{0, \frac{1}{2}, 1\}$ such that:

i. V_M dictates that conjunction ($\wedge$), disjunction ($\vee$), negation ($\neg$), universal ($\forall$) and existential ($\exists$) quantification behave according to the Strong Kleene schema[1].

ii. V_M respects ground model M: $\forall \sigma \in L : V_M(\sigma) = \mathcal{C}_M(\sigma)$

iii. V_M satisfies the identity of truth: $\forall \sigma \in L_T : V_M(T(\overline{\sigma})) = V_M(\sigma)$

Kripke [33] showed that there are various fixed point valuations relative to a (fixed) ground model M. We will use $\mathbf{FP}(L_T, M)$ to denote the class of all fixed point valuations of L_T relative to ground model M, whereas $\mathbf{FP}(L_T)$ will denote the class of all fixed point valuations. That is:

$$V \in \mathbf{FP}(L_T) \Leftrightarrow V \in \mathbf{FP}(L_T, M) \text{ for some ground model } M$$

There are various ways to interpret (the range of) a fixed point valuation. As announced, we will do so in terms of strict and tolerant assertions and denials. With respect to a fixed point valuation $V : Sen(L_T) \to \{0, \frac{1}{2}, 1\}$, the strict-tolerant slang is to be used as follows. Sentences that are valuated as 1 are *strictly assertible*, sentences that are valuated as 0 are *strictly deniable* and sentences that are valuated as $\frac{1}{2}$ are *neither* strictly assertible nor strictly deniable. Sentences that receive a value in $\{1, \frac{1}{2}\}$ are *tolerantly assertible*, whereas those that receive a value in $\{0, \frac{1}{2}\}$ are *tolerantly deniable*. Indeed, sentences that are valuated as $\frac{1}{2}$ are neither strictly assertible nor deniable, but, at the same time, both tolerantly assertible and deniable.

Exploiting the strict-tolerant terminology, we see that the fixed point equivalent of (7.1) can be paraphrased as "whenever all premisses are *strictly* assertible, some conclusion is *strictly* assertible". In line with this paraphrase, we say that the fixed point equivalent of (7.1) is the *strict-strict* consequence relation, denoted $\models^{ss}$. Similarly, we can define the fixed point equivalents of (7.2), (7.3) and (7.4), which are denoted as $\models^{tt}$, $\models^{st}$ and $\models^{ts}$ respectively[2]. Here are the four basic fixed point consequence relations:

[1] We use ($\rightarrow$) to express material implication, defined as usual.

[2] The $\models^{ij}$ notation is due to [11].

- $\Gamma \models^{ss} \Delta$ iff for every $V \in \mathbf{FP}(L_T)$:

$$\forall \alpha \in \Gamma : V(\alpha) = 1 \Rightarrow \exists \beta \in \Delta : V(\beta) = 1$$

"All premisses *s*trictly assertible ⇒ some conclusion *s*trictly assertible".

- $\Gamma \models^{tt} \Delta$ iff for every $V \in \mathbf{FP}(L_T)$:

$$\forall \alpha \in \Gamma : V(\alpha) \in \{1, \tfrac{1}{2}\} \Rightarrow \exists \beta \in \Delta : V(\beta) \in \{1, \tfrac{1}{2}\}$$

"All premisses *t*olerantly assertible ⇒ some conclusion *t*olerantly assertible".

- $\Gamma \models^{st} \Delta$ iff for every $V \in \mathbf{FP}(L_T)$:

$$\forall \alpha \in \Gamma : V(\alpha) = 1 \Rightarrow \exists \beta \in \Delta : V(\beta) \in \{1, \tfrac{1}{2}\}$$

"All premisses *s*trictly assertible ⇒ some conclusion *t*olerantly assertible".

- $\Gamma \models^{ts} \Delta$ iff for every $V \in \mathbf{FP}(L_T)$:

$$\forall \alpha \in \Gamma : V(\alpha) \in \{1, \tfrac{1}{2}\} \Rightarrow \exists \beta \in \Delta : V(\beta) = 1$$

"All premisses *t*olerantly assertible ⇒ some conclusion *s*trictly assertible".

$\models^{ss}$ captures the intuitive principle that whenever one asserts all the premisses of a valid argument, one must also assert its conclusion, while $\models^{tt}$ captures the principle that whenever one denies the conclusion of a valid argument, one must also deny one of its premisses. However, although they capture intuitive principles, both $\models^{ss}$ and $\models^{tt}$ give rise to undesirable, non-classical, behavior of material implication ($\rightarrow$). For, observe that:

$$\not\models^{ss} \alpha \rightarrow \alpha, \qquad \alpha, \alpha \rightarrow \beta \not\models^{tt} \beta$$

Hence, $\models^{ss}$ violates *identity* while $\models^{tt}$ violates *material modus ponens*. To put some flesh on the bones, we give two concrete examples of the violations of these classical principles. To see that $\models^{ss}$ violates identity, let:

$$\alpha := \lambda \approx [\neg T(\lambda)] \wedge \neg T(\lambda)$$

As identity behaves classically, $\lambda \approx [\neg T(\lambda)]$ is, in any fixed point valuation, either valuated as 0 or 1. Whenever it is valuated as 1, $\neg T(\lambda)$ is a *Liar* and hence it must be valuated as $\frac{1}{2}$. In these valuations then, α also valuates as $\frac{1}{2}$, and hence so does $\alpha \rightarrow \alpha$. This establishes that $\models^{ss}$ violates identity. To see that $\models^{tt}$ violates material modus ponens, let α be as before and let:

$$\beta := \lambda \not\approx [\neg T(\lambda)]$$

Again, let V be a fixed point valuation in which $\lambda \approx [\neg T(\lambda)]$ is valuated as 1. Observe that $V(\alpha) = \frac{1}{2}$, $V(\alpha \rightarrow \beta) = \frac{1}{2}$ and $V(\beta) = 0$ and that these three observations jointly establish that $\models^{tt}$ violates material modus ponens. In contrast, we have that:

$$\models^{st} \alpha \rightarrow \alpha, \qquad \alpha, \alpha \rightarrow \beta \models^{st} \beta, \tag{7.5}$$

as the reader may verify by inspecting the "truth tables". Hence, identity and material modus ponens are both *st*-valid. In fact, as shown by [45], this observation is an instance of:

Fact 3 *Any argument form that is classically valid is* $\models^{st}$*-valid.*

Fact 3 certainly is a nice property for a consequence relation to have, and it may provide a reason for preferring $\models^{st}$ over $\models^{ss}$ and $\models^{tt}$. However, with the semantic paradoxes around, any "good" property of a fixed point consequence relation comes at a price; the relation must also give up some intuitively plausible semantic principles. The price that $\models^{st}$ has to pay for Fact 3 is:

Fact 4 $\models^{st}$ *is non-transitive:* $\alpha \models^{st} \beta \;\&\; \beta \models^{st} \gamma \not\Rightarrow \alpha \models^{st} \gamma$

To illustrate Fact 4, observe that:

$$\lambda \approx [\neg T(\lambda)] \models^{st} \neg T(\lambda) \tag{7.6}$$

To see that (7.6) holds, note that when $\lambda \approx [\neg T(\lambda)]$ valuates as 1, $\neg T(\lambda)$ is a Liar, which must be valuated as $\frac{1}{2}$. Hence, we have (7.6). Further, observe that:

$$\neg T(\lambda) \models^{st} \lambda \not\approx [\neg T(\lambda)] \tag{7.7}$$

Now, whenever $\neg T(\lambda)$ is valuated as 1, $\lambda \not\approx [\neg T(\lambda)]$ must also be valuated as 1. For, suppose that $\neg T(\lambda)$ is valuated as 1 and that $\lambda \not\approx [\neg T(\lambda)]$ is not valuated as 1. Then, as identity behaves classically (and given the behavior of $\neg$) $\lambda \approx [\neg T(\lambda)]$ must be valuated as 1. But then $\neg T(\lambda)$ is a Liar, which is valuated as $\frac{1}{2}$ in any fixed point valuation. This establishes (7.7). Now, we clearly have that:

$$\lambda \approx [\neg T(\lambda)] \not\models^{st} \lambda \not\approx [\neg T(\lambda)] \tag{7.8}$$

Thus, (7.6), (7.7) and (7.8) jointly illustrate the non-transitivity of $\models^{st}$.

Our discussion of $\models^{ss}$, $\models^{tt}$ and $\models^{st}$ leaves us with $\models^{ts}$. Here we can be short: although the relation is of formal interest—leaving it out violates symmetry considerations—it is of little interest as a genuine consequence relation. For instance, observe that:

$$T(t) \not\models^{ts} T(t) \tag{7.9}$$

To see that (7.9) holds, note that there are no restrictions on how a fixed point valuation treats sentences of form $T(t)$, with t a non quotational constant: there are fixed point valuations[3] in which $T(t)$ valuates as 0, $\frac{1}{2}$ and 1. Any fixed point in which $T(t)$ valuates as $\frac{1}{2}$ establishes (7.9). As the reader may verify for himself, $\models^{ts}$ has a lot more undesirable properties.

7.2.2 The Strict-Tolerant calculus

Now that we have introduced the four fixed point relations, we are ready to sketch how we will capture them syntactically. We will develop a signed tableau calculus, whose four signs (A^s, D^s, A^t and D^t) correspond to the four assertoric acts associated with the strict-tolerant interpretation in the expected manner. Formally, one may think of $A^s_\sigma, D^s_\sigma, A^t_\sigma$ and D^t_σ as expressing that the semantic value of sentence σ is contained in, respectively, $\{1\}, \{0\}, \{1, \frac{1}{2}\}$ and $\{0, \frac{1}{2}\}$. Our calculus, whose tableau expansion rules and closure conditions are discussed

[3] As Michael Kremer [29] observes, the consequence relations based on the class of all fixed point valuations according to which all truth ascriptions to non-sentences are valuated as 0 or $\frac{1}{2}$, are not compact.

below, is called the *strict-tolerant calculus*.

A signed—with A^s, D^s, A^t or D^t—sentence will also be called an *assertoric sentence*. For sake of definiteness, we display the tableau rules for the quantifier free part of L_T below, where $i \in \{s,t\}$.

$$\frac{A^i_{\neg\alpha}}{D^i_\alpha} \quad \frac{D^i_{\neg\alpha}}{A^i_\alpha} \qquad \frac{A^i_{\alpha\vee\beta}}{A^i_\alpha \mid A^i_\beta} \quad \frac{D^i_{\alpha\vee\beta}}{D^i_\alpha, D^i_\beta} \qquad \frac{A^i_{\alpha\wedge\beta}}{A^i_\alpha, A^i_\beta} \quad \frac{D^i_{\alpha\wedge\beta}}{D^i_\alpha \mid D^i_\beta}$$

$$\frac{A^i_{T([\alpha])}}{A^i_\alpha} \qquad \frac{D^i_{T([\alpha])}}{D^i_\alpha}$$

$$\frac{A^i_{a\approx b}}{A^i_{b\approx a}} \quad \frac{A^s_{a\approx b}}{A^t_{a\approx b}} \quad \frac{A^t_{a\approx b}}{A^s_{a\approx b}} \qquad \frac{A^i_{a\approx b}, A^i_{\phi(a)}}{A^i_{\phi(a/b)}} \quad \frac{A^i_{a\approx b}, D^i_{\phi(a)}}{D^i_{\phi(a/b)}}$$

Looking at the tableau rules, we easily see that these rules are *valid* in every fixed point valuation V. For instance, the rule $A^s_\neg$ is valid as:

$$\forall V \in \mathbf{FP}(L_T) : V(\neg\alpha) = 1 \Leftrightarrow V(\alpha) = 0$$

While the fact that $A^t_\neg$ is valid means that:

$$\forall V \in \mathbf{FP}(L_T) : V(\neg\alpha) \in \{1, \tfrac{1}{2}\} \Leftrightarrow V(\alpha) \in \{0, \tfrac{1}{2}\}$$

Similarly, the reader can verify that all other tableau rules are valid in this sense. As any strict tableau rule has a tolerant counterpart and vice versa, the validity of the tableau rules indicates that strict and tolerant assertion and denial are *governed by the same assertoric rules*. However, although $\neg\alpha$ is strictly / tolerantly assertible just in case α is strictly / tolerantly deniable, it may very well be that $\neg\alpha$ is tolerantly assertible without being strictly assertible, as the Liar testifies. This distinction is explained by the *norm* that governs the (strict /tolerant) assertoric actions, which is formally represented by the *closure conditions* of our tableau calculus. A set of assertoric sentences S is *closed* just in case:

1. For some arbitrary sentence σ: $\{A^s_\sigma, D^s_\sigma\} \subseteq S$
2. For some truth-free[4] sentence σ: $\{A^t_\sigma, D^t_\sigma\} \subseteq S$
3. For some arbitrary sentence σ: $\{A^s_\sigma, D^t_\sigma\} \subseteq S$
4. For some arbitrary sentence σ: $\{A^t_\sigma, D^s_\sigma\} \subseteq S$
5. For some constant c: $D^i_{c\approx c} \in S$ $\qquad i \in \{s,t\}$

[4]In a truth-free sentence, the truth predicate may only occur within a quotational constant; $P([T(c)])$ is a truth-free sentence but $T([P(c)])$ is not.

6. For *distinct* sentences α, β: $A^i_{[\alpha]\approx[\beta]} \in S$ $\quad i \in \{s,t\}$

The first closure condition states that it is never allowed to strictly assert and deny the same sentence. As we saw above, it is allowed to both tolerantly assert and deny some sentences, as in the case of the Liar. However, the second closure condition tells us that is not allowed to both tolerantly assert and deny truth-free sentences, which may be thought of as "unproblematic declarative sentences describing non-semantic states of affairs". The third and fourth closure condition state that strictly (tolerantly) asserting σ rules out tolerantly (strictly) denying σ and vice versa. The fifth closure condition indicates that it is never allowed to deny (strictly or tolerantly) trivial identity statements and the sixth closure condition indicates that it is never allowed to assert (strictly or tolerantly) that two distinct sentences are identical.

A notion that plays a crucial role in (tableau based) soundness and completeness proofs for classical logic is that of *satisfiability*. In our, to be given, soundness and completeness proofs pertaining to the fixed point consequence relations, a similar role is played by the notion of *fixed point satisfiability*. A set of assertoric sentences S is said to be *fixed point satisfiable* just in case there exists a fixed point valuation V such that:

$$A^s_\sigma \in S \Rightarrow V(\sigma) = 1, \qquad D^s_\sigma \in S \Rightarrow V(\sigma) = 0, \tag{7.10}$$

$$A^t_\sigma \in S \Rightarrow V(\sigma) \in \{1, \tfrac{1}{2}\}, \qquad D^t_\sigma \in S \Rightarrow V(\sigma) \in \{0, \tfrac{1}{2}\}. \tag{7.11}$$

The notion of fixed point satisfiability allows us to reformulate the fixed point consequence relations in terms of sets of assertoric sentences. Observe that per definition:

$$\Gamma \models^{st} \Delta \Leftrightarrow \{A^s_\alpha \mid \alpha \in \Gamma\} \cup \{D^s_\beta \mid \beta \in \Delta\} \text{ is not fixed point satisfiable.} \tag{7.12}$$

$$\Gamma \models^{ss} \Delta \Leftrightarrow \{A^s_\alpha \mid \alpha \in \Gamma\} \cup \{D^t_\beta \mid \beta \in \Delta\} \text{ is not fixed point satisfiable.} \tag{7.13}$$

$$\Gamma \models^{tt} \Delta \Leftrightarrow \{A^t_\alpha \mid \alpha \in \Gamma\} \cup \{D^s_\beta \mid \beta \in \Delta\} \text{ is not fixed point satisfiable.} \tag{7.14}$$

$$\Gamma \models^{ts} \Delta \Leftrightarrow \{A^t_\alpha \mid \alpha \in \Gamma\} \cup \{D^t_\beta \mid \beta \in \Delta\} \text{ is not fixed point satisfiable.} \tag{7.15}$$

This reformulation motivates the following definition of $\vdash^{st}$, $\vdash^{ss}$, $\vdash^{tt}$ and $\vdash^{ts}$. Say that a set of assertoric sentences S is *expansion closed* just in case there exists a (finite) tableau starting in S which is closed. We let:

- $\Gamma \vdash^{st} \Delta \Leftrightarrow \{A^s_\alpha \mid \alpha \in \Gamma\} \cup \{D^s_\beta \mid \beta \in \Delta\}$ is expansion closed.
- $\Gamma \vdash^{ss} \Delta \Leftrightarrow \{A^s_\alpha \mid \alpha \in \Gamma\} \cup \{D^t_\beta \mid \beta \in \Delta\}$ is expansion closed.
- $\Gamma \vdash^{tt} \Delta \Leftrightarrow \{A^t_\alpha \mid \alpha \in \Gamma\} \cup \{D^s_\beta \mid \beta \in \Delta\}$ is expansion closed.
- $\Gamma \vdash^{ts} \Delta \Leftrightarrow \{A^t_\alpha \mid \alpha \in \Gamma\} \cup \{D^t_\beta \mid \beta \in \Delta\}$ is expansion closed.

By following the structure of tableau-based soundness and completeness proofs for classical logic (see, e.g., [50]), we will prove, in Section 2, the main theorem of this paper:

Theorem: With $i, j \in \{s,t\}$: $\Gamma \vdash^{ij} \Delta \Leftrightarrow \Gamma \models^{ij} \Delta$

That is, each of our syntactic consequence relations is sound and complete with respect to its semantic counterpart. Further, Section 2 shows how to define the *classical calculus* as a sub calculus of the strict-tolerant calculus and, by exploiting the classical calculus, we give an instructive proof of Fact 3.

7.2.3 Assertoric Semantics

In Section 3, we present a semantic version of our strict-tolerant calculus, called *assertoric semantics*. Assertoric semantics answers the *actuality question*: given a fixed (actual) ground model M, which sentences are strictly / tolerantly assertible and / or deniable in M? To answer the actuality question, assertoric semantics augments the closure conditions of the strict-tolerant calculus with closure conditions that model assertoric norms stemming from the ground model M under consideration. Further, assertoric semantics slightly modifies the rules of the strict-tolerant calculus in a straightforward manner to account for the fact that we are dealing with a fixed ground model M. Given a ground model M, assertoric semantics returns two L_T valuation functions: the *tolerant valuation function* $\mathcal{V}^t_M$ and the *strict valuation function* $\mathcal{V}^s_M$. The functions $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$ inform us, respectively, about the tolerant and the strict assertoric status of the L_T sentences relative to M. It turns out that, as we will prove, $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$ are familiar functions. $\mathcal{V}^t_M$ is equivalent to the (Strong Kleene) minimal fixed point valuation over M, whereas $\mathcal{V}^s_M$ is equivalent to the function that Kripke [33] (implicitly) defined by quantifying over all fixed point valuations. For instance, $\mathcal{V}^s_M$ will valuate the Liar as $(0, 0)$, indicating that it is forbidden to strictly assert and to strictly deny the Liar (in M). According to Kripke, the Liar is *paradoxical*, as there is no fixed point valuation (over M) that valuates it as 1 and, also there is no fixed point valuation (over M) that valuates the Liar as 0. The function $\mathcal{V}^t_M$, however, will valuate the Liar as $(1, 1)$ indicating that the Liar is both tolerantly assertible and deniable (in M). In terms of the minimal fixed point, this corresponds with the Liar being valuated as $\frac{1}{2}$.

7.2.4 STCT

In [46], Ripley advocates a new approach to truth and semantics, which heavily relies on the strict-tolerant distinction. Ripley's conception of truth will be called the *Strict Tolerant Conception of Truth* (STCT). Here are the most distinguishing features of STCT.

1. STCT advocates an inferentialist, *bilateralist* theory of meaning.
   In a nutshell, *inferentialism* is the view that meanings are to be explained in terms of which inferences are valid, while *bilateralism* is a species of interentialism (defended by e.g., Rumfitt [48]) according to which the validity of inferences is to explained in terms of conditions on *assertion* and *denial*. Most notably, bilateralism acknowledges two primitive assertoric speech acts: assertion and denial. This distinguishes bilateralism from *unilateralism* (e.g., Geach [19]), which holds that assertion is the only primitive assertoric speech act and according to which a denial is to be understood as the assertion of a negation. Bilateralism reverses the order of explanation; negation is to be understood in terms of denial.

2. STCT advocates $\models^{st}$ as the norm according to which L_T inferences should be valuated as (in)correct. More precisely, the bilateralist theory of meaning for L_T is realized in the form of a 2-sided sequent calculus that is sound and complete with respect to $\models^{st}$. Indeed, being an inferentialist position, the sequent calculus (characterization of $\models^{st}$) is fundamental for STCT.

3. STCT acknowledges four distinct (strict-tolerant) assertoric actions. The commitment to *four* distinct assertoric actions (rather than one) is, *prima facie*, an unattractive feature of STCT. However, Ripley [46] has argued that this need not be the case, for the strict-tolerant distinction is not a *primitive* one. Rather, the strict can be understood in terms of the tolerant or vice versa.

As the non-transitivity of $\models^{st}$ is, *prima facie*, an undesirable property, an STCT proponent has to argue that, here, first looks are deceiving. For such an argument, see [46]. In this paper, we will simply accept the non-transitivity of $\models^{st}$ as a mathematical fact.

Our remarks on STCT will mainly concern distinguishing feature 3. By exploiting the strict-tolerant calculus and assertoric semantics, we will argue that STCT is committed to acknowledging four *primitive* assertoric actions. Then, we will consider whether our conclusion has consequences for STCT's self-declared commitment to bilateralism.

7.3 The Strict-Tolerant Calculus

The language L_T^U with parameters in U. In order to set up our soundness and completeness proofs, we will work with the language L_T^U, which is obtained by extending L_T with a set of constant symbols $U = \{u_1, u_2, \ldots\}$ disjoint from $Con(L_T)$. Elements of U are called *parameters*. The language L_T^U is obtained by adding the parameters of U to L_T and by *closing off under the formation of quotational constants*. For example, although $u_1 \approx u_2$ is not a sentence of L_T, it is a sentence of L_T^U and, accordingly, $[u_1 \approx u_2]$ is a quotational constant of L_T^U. We will use $Con(L_T^U)$ to denote the set of all (including parameters) constant symbols of L_T^U. The notion of a ground model and fixed point valuation of L_T^U are defined similar to the notions of a ground model and fixed point valuation of L_T. We will use $\mathbf{FP}(L_T^U)$ to denote the class of all fixed point valuations of L_T^U.

The Strict-Tolerant Calculus. The *tableau expansion* rules of the strict-tolerant calculus will manipulate signed L_T^U sentences. The expansion rules for the quantifier-free part were displayed in the previous section. Here are the expansion rules for the quantifiers, where $i \in \{s, t\}$:

$$\frac{A^i_{\forall x\phi(x)}}{A^i_{\phi(x/c)}} \qquad \frac{D^i_{\forall x\phi(x)}}{D^i_{\phi(x/u)}}\ u \text{ fresh} \qquad\qquad \frac{A^i_{\exists x\phi(x)}}{A^i_{\phi(x/u)}}\ u \text{ fresh} \qquad \frac{D^i_{\exists x\phi(x)}}{D^i_{\phi(x/c)}}$$

In the rules $A^i_\forall$ and $D^i_\exists$, c is an arbitrary element of $Con(L_T^U)$. Likewise, in the rules for identity that were displayed before, a and b are arbitrary elements of

$Con(L_T^U)$. The slogan 'u is fresh' that accompanies the rules $A_\exists^i$ and $D_\forall^i$ means that the *parameter* $u \in U$ does not occur (in some sentence) on the path that is extended when applying the $A_\exists^i$ or $D_\forall^i$ rule.

The *closure rules* of the strict-tolerant calculus were displayed in the previous section. Note that by an *arbitrary sentence* (cf. closure conditions 1, 3, 4 and 6) we mean a sentence of L_T^U, by a *truth-free sentence* (cf. closure condition 2) a sentence of $L^U = L_T^U - \{T\}$ and by a *constant* (cf. closure condition 5) we mean an element of $Con(L_T^U)$. Sets of signed L_T^U sentences that are not closed are called *open*.

The relations $\vdash^{st}$, $\vdash^{ss}$, $\vdash^{tt}$ and $\vdash^{ts}$. These relations were already defined in the previous section by employing the notion of the *expansion closure* of a set of *assertoric sentences*. Here, an assertoric sentence is a signed sentence of L_T^U. To be sure, with S a (finite or infinite) set of assertoric sentences, we say that S is *expansion closed* just in case there is a tableau starting with some *finite* $S' \subseteq S$ which is closed.

Fixed point satisfiability and fixed point$_U$ satisfiability With S a set of signed sentences of L_T, we say that S is *fixed point satisfiable* just in case there is some $V \in \mathbf{FP}(L_T)$ such that conditions (7.10) and (7.11) hold. Similarly, with S a set of signed sentences of L_T^U, we say that S is *fixed point$_U$ satisfiable* just in case there is some $V \in \mathbf{FP}(L_T^U)$ such that conditions (7.10) and (7.11) hold. With S a set of signed sentences of L_T, we have, as $Sen(L_T) \subseteq Sen(L_T^U)$, that both our satisfiability notions are applicable to S. The following lemma tells us that it does not matter which of the two we apply.

Lemma 7.1 With $S \subseteq Sen(L_T)$: S is fixed point satisfiable iff S is fixed point$_U$ satisfiable.
Proof: Left to the reader. □

Preservation of fixed point$_U$ satisfiability. A tableau $\mathbf{T}$ is said to be *fixed point$_U$ satisfiable* just in case one of its paths is fixed point$_U$ satisfiable. The following lemma tells us that the tableau rules preserve fixed point$_U$ satisfiability.

Lemma 7.2 Preservation of fixed point$_U$ satisfiability
Let $\mathbf{T}$ and $\mathbf{T}'$ be tableaux such that $\mathbf{T}'$ is an *immediate extension* of $\mathbf{T}$, i.e., $\mathbf{T}'$ is obtained from $\mathbf{T}$ by applying a tableau rule to a path of $\mathbf{T}$. If $\mathbf{T}$ is fixed point$_U$ satisfiable, so is $\mathbf{T}'$.
Proof: A case by case inspection of the tableau rules which can safely be left to the reader. □

Together with the observation that no closed set of assertoric sentences is fixed point$_U$ satisfiable, Lemma 7.2 easily implies the *soundless lemma*.

Lemma 7.3 Soundness Lemma
Let S be a finite set of signed sentences of L_T^U. If there exists a closed tableau starting with S, then S is not fixed point$_U$ satisfiable.
Proof: Let $\mathbf{T}$ be a closed tableau starting with S. Thus, there is a finite sequence of tableaux $\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_n = \mathbf{T}$ such that $\mathbf{T}_0 = S$ and such that $\mathbf{T}_{i+1}$ is an immediate extension of $\mathbf{T}_i$. Suppose S is fixed point$_U$ satisfiable. Then $\mathbf{T}_0$ is

fixed point$_U$ satisfiable per definition and, as the tableau rules preserve fixed point$_U$ satisfiability (cf. Lemma 7.2), so is $\mathbf{T}_n = \mathbf{T}$. However, as $\mathbf{T}$ is closed, this is impossible. □

Our lemma's culminate in the soundness theorem:

Theorem 7.1 Soundness: $i, j \in \{s,t\}, \Gamma, \Delta \subseteq Sen(L_T)$**:** $\Gamma \vdash^{ij} \Delta \Rightarrow \Gamma \models^{ij} \Delta$
Proof: From the definition of $\models^{ij}$ in terms of fixed point satisfiability (as in (7.12), (7.13), (7.14) and (7.15)), Lemma 7.1 and Lemma 7.3. □

Downwards saturated sets. A set of assertoric sentences S is called *downwards saturated* just in case S satisfies the following conditions, where $i \in \{s,t\}$:

1. $A^i_{\neg\alpha} \in S \Rightarrow D^i_\alpha \in S$, $\quad D^i_{\neg\alpha} \in S \Rightarrow A^i_\alpha \in S$
2. $A^i_{\alpha\vee\beta} \in S \Rightarrow A^i_\alpha \in S$ or $A^i_\beta \in S$, $\quad D^i_{\alpha\vee\beta} \in S \Rightarrow D^i_\alpha \in S$ & $D^i_\beta \in S$
3. $A^i_{\alpha\wedge\beta} \in S \Rightarrow A^i_\alpha \in S$ & $A^i_\beta \in S$, $\quad D^i_{\alpha\wedge\beta} \in S \Rightarrow D^i_\alpha \in S$ or $D^i_\beta \in S$
4. $A^i_{T([\alpha])} \in S \Rightarrow A^i_\alpha \in S$, $\quad D^i_{T([\alpha])} \in S \Rightarrow D^i_\alpha \in S$
5. $A^i_{\forall x\phi(x)} \left(D^i_{\exists x\phi(x)}\right) \in S \Rightarrow A^i_{\phi(x/c)} \left(D^i_{\phi(x/c)}\right) \in S$ for all $c \in Con(L^U_T)$
6. $A^i_{\exists x\phi(x)} \left(D^i_{\forall x\phi(x)}\right) \in S \Rightarrow A^i_{\phi(x/c)} \left(D^i_{\phi(x/c)}\right) \in S$ for some $c \in Con(L^U_T)$
7. $A^i_{a\approx b} \in S \Rightarrow A^i_{b\approx a} \in S$

8a. $A^s_{a\approx b} \in S \Rightarrow A^t_{a\approx b} \in S$, $\quad$ 8b. $A^t_{a\approx b} \in S \Rightarrow A^s_{a\approx b} \in S$

Sub: $A^i_{a\approx b}$ & $A^i_{\phi(a)} \left(A^i_{a\approx b}\ \&\ D^i_{\phi(a)}\right) \in S \Rightarrow A^i_{\phi(a/b)} \left(D^i_{\phi(a/b)}\right) \in S$

Let S be a set of *strict assertoric sentences*, i.e., $X^i_\sigma \in S \Rightarrow i = s$. We say that S is *strictly downwards saturated* (also, *downwards$_s$ saturated*) just in case S satisfies (the strict version[5] of) the conditions 1-7 and Sub.

Upwards saturated sets. Each downwards saturation condition C that is labeled with 1-8 has form $P \Rightarrow Q$ and the *upwards saturation condition associated with such a* C is given by $P \Leftarrow Q$. A set of assertoric sentences S is called *upwards saturated* just in case S satisfies the upwards saturation conditions associated with the downwards saturation conditions labeled by 1-8, together with Sub (that is the reading of Sub is not reversed). A set of strict assertoric sentences S is called *strictly upwards saturated* (also, *upwards$_s$ saturated*) just in case S satisfies the upwards saturation conditions associated with the downwards saturation conditions labeled by 1-7, together with Sub (that is the reading of Sub is not reversed).

Upwards closure. It can be shown, by arguments familiar from, e.g., [16], that for each set of assertoric sentences S there is a smallest upwards saturated set extending S. We call this set the *upwards closure of* S. Similarly, each downwards$_s$ saturated set of strict assertoric sentences has a smallest upwards$_s$ saturated set extending it, which will be called the *strict upwards closure of* S.

The following theorem will constitute the heart of our completeness proof.

[5]The tolerant versions are trivially satisfied as S is a set of strict assertoric sentences.

Theorem 7.2 Open downwards saturated sets are fixed point$_U$ satisfiable

Proof: Let S be an open, downwards saturated set of signed sentences of L_T^U. Define an equivalence relation, $\equiv$ on $Con(L_T^U)$ as follows:

$$\equiv \; = \; \{(a,b) \mid A^s_{a \approx b} \in S\} \cup \{(a,a) \mid a \in Con(L_T^U)\}$$

Now $\equiv$ is indeed an equivalence relation on $Con(L_T^U)$. Reflexivity follows per definition and transitivity and symmetry follow from the downwards saturation conditions (including Sub) pertaining to identity. For any $c \in Con(L_T^U)$, we let $\langle c \rangle$ be the equivalence class (induced by $\equiv$) containing c. We now define a ground model $M = (D, I)$ as follows. Our domain D consists of all sentences of L_T^U together with all equivalence classes that do not contain quote names:

$$D = Sen(L_T^U) \cup \{\langle c \rangle \mid \text{for no } \sigma \in Sen(L_T^U) : [\sigma] \in \langle c \rangle\}$$

The interpretation function I is defined as follows. For any $c \in Con(L_T^U)$:

$$I(c) = \begin{cases} \langle c \rangle; & \text{for no } \sigma \in Sen(L_T^U) : [\sigma] \in \langle c \rangle \\ \sigma; & [\sigma] \in \langle c \rangle \end{cases}$$

Note that the second clause is well-defined; given that S is open, we never have that $A^i_{[\alpha] \approx [\beta]} \in S$ for distinct $\alpha, \beta \in Sen(L_T^U)$. Hence, there is at most one $[\sigma] \in Sen(L_T^U)$ occurring in an equivalence class $\langle c \rangle$. The extension of $\approx$ is dictated by the fact that $\approx$ is the identity relation on D, while the extension of an n-ary predicate $P \neq \approx$ of L^U is given as follows:

$$I(P) = \{(I(t_1), \ldots, I(t_n)) \mid A^s_{P(t_1, \ldots t_n)} \in S \text{ or } A^t_{P(t_1, \ldots t_n)} \in S\}$$

The ground model M defined as such gives rise to a classical valuation $\mathcal{C}_M : Sen(L^U) \to \{1, 0\}$. We use this valuation to define the set w_s, which may be called (the strict version of) the *world*. Below, $At(L^U)$ is the set of all atomic sentences of L^U.

$$w_s = \{A^s_\sigma \mid \sigma \in At(L^U), \mathcal{C}_M(\sigma) = 1\} \cup \{D^s_\sigma \mid \sigma \in At(L^U), \mathcal{C}_M(\sigma) = 0\}$$

We will now show how we can construct a fixed point valuation based on M. To do so, we first define S^s as the set consisting of all strict assertoric sentences occurring in S respectively. Thus:

$$S^s = \{X^s_\sigma \mid X^s_\sigma \in S, X \in \{A, D\}\}$$

Next, we extend S^s to the set S^s_w by adding w_s to it:

$$S^s_w = S^s \cup w_s$$

Note that, although S^s_w contains only strict assertoric sentences, it contains ("encoded in strict terms") information about the tolerant assertoric sentences of $At(L^U)$ that occurred in S. Also, observe that S^s_w is open and downwards$_s$ saturated. Let $S^{s\Uparrow}_w$ be the upwards$_s$ closure of S^s_w and define the function V_M as follows:

$$V_M(\sigma) = \begin{cases} 1 & A^s_\sigma \in S^{s\Uparrow}_w \\ \frac{1}{2} & \{A^s_\sigma, D^s_\sigma\} \cap S^{s\Uparrow}_w = \varnothing \\ 0, & D^s_\sigma \in S^{s\Uparrow}_w \end{cases}$$

Indeed, V_M is well-defined: as S^s_w is open and downwards$_s$ saturated, $S^{s\Uparrow}_w$ is open, downwards$_s$ saturated and upwards$_s$ saturated (by arguments familiar from [16]). Hence, as $S^{s\Uparrow}_w$ is open we never have that both A_σ and D_σ occur in $S^{s\Uparrow}_w$, so that V_M is well-defined. Further, from the fact that $S^{s\Uparrow}_w$ is both downwards$_s$ and upwards$_s$ saturated, it follows that V_M is a Strong Kleene valuation of L^U_T which respects the identity of truth. Moreover, as S^s_w contains either A_σ or D_σ for each atomic L^U sentence in accordance with $\mathcal{C}_M$, it follows that $S^{s\Uparrow}_w$ contains either A_σ or D_σ for each L^U sentence in accordance with $\mathcal{C}_M$. Hence, V_M respects the ground model M. Thus, $V_M \in \mathbf{FP}(L^U_T)$. Finally, as $S^s \subseteq S^{s\Uparrow}_w$, we have that:

$$A^s_\sigma \in S^s \Rightarrow V_M(\sigma) = 1, \qquad D^s_\sigma \in S^s \Rightarrow V_M(\sigma) = 0 \tag{7.16}$$

As S^s contains the strict part of S, it suffices, in order to show that S is fixed point satisfiable, to show that:

$$A^t_\sigma \in S \Rightarrow V_M(\sigma) \in \{1, \tfrac{1}{2}\}, \qquad D^t_\sigma \in S \Rightarrow V_M(\sigma) \in \{0, \tfrac{1}{2}\} \tag{7.17}$$

We will show that $A^t_\sigma \in S \Rightarrow V_M(\sigma) \in \{1, \frac{1}{2}\}$; the argument for the D^t case is completely dual. We give a reductio argument. Suppose that $A^t_\sigma \in S$ and $V_M(\sigma) = 0$ and observe that the latter is equivalent to $D^s_\sigma \in S^{s\Uparrow}_w$. We will show that $A^t_\sigma \in S$ implies that $D^s_\sigma \notin S^{s\Uparrow}_w$ by transfinite induction. In order to do so, we define the sequence $\{S^\rho\}_{\rho \in On}$ consisting of sets of assertoric sentences.

- $\rho = 0$: $S^\rho = S^s_w$
- $\rho = \upsilon + 1$: S^ρ is obtained by *extending* S^υ by applying the upwards$_s$ rules (including Sub) that are allowed on the basis of S^υ. For instance, if A^s_α and $A^s_\beta \in S^\upsilon$ then A^s_α, $A^s_\beta, A^s_{\alpha\wedge\beta} \in S^\rho$. Similarly for all other upwards$_s$ rules.
- ρ is a limit ordinal: $S^\rho = \bigcup_{\upsilon<\rho} S^\upsilon$.

Clearly, $\{S^\rho\}_{\rho \in On}$ is an increasing sequence and has a fixed point which is equal to $S^{s\Uparrow}_w$. Hence, it suffices to show that for any $\rho \in On$:

$$A^t_\sigma \in S \Rightarrow D^s_\sigma \notin S^\rho \tag{7.18}$$

Induction basis. From the fact that S is open, it follows that $A^t_\sigma \in S \Rightarrow D^s_\sigma \notin S^s$. From the way in which we defined the world, we get that $A^t_\sigma \in S \Rightarrow D^s_\sigma \notin S^s_w = S^0$.

Successor step. Suppose that (7.18) holds for ρ. We show that (7.18) also holds for $\rho + 1$. To do so, we show that extending S^ρ by applying an upwards$_s$ rule to S^ρ results in a set for which (7.18) still holds. We illustrate two cases, leaving the rest for the reader. In both cases, we argue by contraposition. First case: suppose that $A^t_{\neg\alpha} \in S$ and $D^s_{\neg\alpha} \in S^{\rho+1}$. From the first it follows, as S

is downwards$_s$ closed, that $D^t_\alpha \in S$. From the second, it follows that $A^s_\alpha \in S^\rho$ per definition of $S^{\rho+1}$. Hence, we have that $A^t_\alpha \in S$ and $D^s_\alpha \in S^\rho$, which gives a contradiction with the assumption that (7.18) holds for ρ. Second case: suppose that $A^t_{\alpha\vee\beta} \in S$ and $D^s_{\alpha\vee\beta} \in S^{\rho+1}$. From the latter, it follows that $D^s_\alpha \in S^\rho$ and $D^s_\beta \in S^\rho$. As S is downwards saturated, $A^t_{\alpha\vee\beta} \in S$ implies that $A^t_\alpha \in S$ or $D^t_\beta \in S$. So we have that: $\big(A^t_\alpha \in S \;\&\; D^s_\alpha \in S^\rho\big)$ or $\big(A^t_\beta \in S \;\&\; D^s_\beta \in S^\rho\big)$. Either way, we get a contradiction with the assumption that (7.18) holds for ρ.

Limit step. Suppose that ρ is a limit ordinal and that (7.18) holds for all $\upsilon < \rho$. As S^ρ is the union of all S^υ with $\upsilon < \rho$, (7.18) also holds for S^ρ.

So, we have established that (7.18) holds for all ordinals, from which it follows that $A^t_\sigma \in S \Rightarrow D^s_\sigma \notin S^{s\Uparrow}_w$, which is what we had to show. □

To establish completeness, we follow a route familiar from, e.g., [50]. First, we establish *finite completeness*, i.e., we show that, for finite Γ and Δ, with $i, j \in \{s, t\}$: if $\Sigma \vDash^{ij} \Delta$ then $\Sigma \vdash^{ij} \Delta$. Second, we show that $\vDash^{ij}$ is compact. Then, the completeness theorem (pertaining to arbitrary sets of sentences) immediately follows from finite completeness and compactness. Finite completeness follows rather easily from the finite tableau expansion lemma.

Lemma 7.4 Finite tableau extension lemma
Let $\mathbf{T}_0$ be a finite tableau. By applying tableau rules, we can extend $\mathbf{T}_0$ to a (possibly infinite) tableau $\mathbf{T}$ with the following properties: every closed path of $\mathbf{T}$ is finite and every open path of $\mathbf{T}$ is downwards saturated.

Proof: Let $\{t_i\}_{i\in\mathbb{N}}$ be an enumeration of $Con(L^U_T)$. We will start with $\mathbf{T}_0$ and use tableau rules to construct a sequence of finite extensions $\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_2, \ldots$. If some $\mathbf{T}_n$ is closed, then the construction halts, i.e., $\mathbf{T}_m = \mathbf{T}_n$ for all $m \geq n$ and we set $\mathbf{T} = \mathbf{T}_n$. In any case, we set $\mathbf{T} = \bigcup_{i\in\mathbb{N}} \mathbf{T}_i$. Suppose that $\mathbf{T}_n$ is already constructed. Here is how $\mathbf{T}_{n+1}$ is obtained from $\mathbf{T}_n$:

- Define $\mathbf{T}'_n$ as follows. For each open path $\mathcal{P}$ of $\mathbf{T}_n$ and with $j \in \{s, t\}$: when $X^j_\sigma \in \mathcal{P}$ and when X^j_σ is of form $A^j_{\forall x\phi(x)}$ or $D^j_{\exists x\phi(x)}$, extend $\mathcal{P}$ by adding, for each $i \leq n$, respectively, $A^j_{\phi(x/t_i)}$ or $D^j_{\phi(x/t_i)}$. Let $\mathbf{T}'_n$ be the tableau thus obtained.

- Define $\mathbf{T}''_n$ as follows. For each open path $\mathcal{P}$ of $\mathbf{T}'_n$ and with $j \in \{s, t\}$: when $X^j_\sigma \in \mathcal{P}$ and when X^j_σ is of form $X^j_{\neg\alpha}$, $X^j_{\alpha\vee\beta}$, $X^j_{\alpha\wedge\beta}$, $X^j_{T([\alpha])}$, $D^j_{\forall x\phi(x)}$ or $A^j_{\exists x\phi(x)}$, extend $\mathcal{P}$ by applying the appropriate tableau rule. Let $\mathbf{T}''_n$ be the tableau thus obtained.

- Define T_{n+1} as follows. Close off each open path of $\mathbf{T}''_n$ under the tableau rules for identity. T_{n+1} is the tableau thus obtained.

In the construction, a closed path is never extended, so all closed paths are finite. In addition, the construction ensures that each open path of $\mathbf{T}$ is downwards saturated. □

Theorem 7.3 Finite Completeness
Let S be finite set of assertoric sentences. 1) If S is not fixed point satisfiable, then there exists a closed tableau starting with S. 2) Hence, with Γ and Δ finite

set of assertoric sentences and with $i, j \in \{s, t\}$, we have that: if $\Gamma \models^{ij} \Delta$ then $\Gamma \vdash^{ij} \Delta$.

Proof: To establish our first claim, we argue by contraposition. If there does not exists a closed tableau starting with S, the construction from Lemma 7.4 delivers, with $\mathbf{T}_0 = S$, an open downwards saturated path $\mathcal{P}$ such that $S \subseteq \mathcal{P}$. Use $\mathcal{P}$ to construct a fixed point valuation V_M as in Theorem 7.2 and observe that, as $S \subseteq \mathcal{P}$, V_M establishes that S is fixed point satisfiable. This establishes the first claim.

The second claim is a (familiar) consequence of the first claim which we illustrate only for the *st*-variant. Again, we argue by contraposition. If $\Gamma \not\models^{st} \Delta$ then, per definition of $\vdash^{st}$, we have that no tableau starting with $\{A^s_\sigma \mid \sigma \in \Gamma\}$ $\cup$ $\{D^s_\sigma \mid \sigma \in \Gamma\}$ is closed. Hence, from our first claim it follows that $\{A^s_\sigma \mid \sigma \in \Gamma\}$ $\cup$ $\{D^s_\sigma \mid \sigma \in \Gamma\}$ is fixed point satisfiable. Thus, per definition of fixed point satisfiability, there exists a fixed point valuation in which all of Γ is valuated as 1 and in which all of Δ is valuated as 0. Hence, not any fixed valuation which is such that when all of Γ is valuated as 1, some of Δ is valuated as 1 or $\frac{1}{2}$. Hence, $\Gamma \not\models^{st} \Delta$. □

Theorem 7.4 Compactness
Let S be an infinite set of signed sentences of L_T. If each finite subset S is fixed point satisfiable, then S is fixed point satisfiable.

Proof: Enumerate the elements of S, i.e., let $S = \{s_i\}_{i \in \mathbb{N}}$. We will use tableau rules to construct a sequence of finite extensions $\mathbf{T}_0, \mathbf{T}_1, \mathbf{T}_2, \ldots$. Let $\mathbf{T}_0$ be the empty tableau. Suppose $\mathbf{T}_n$ has been constructed. Extend $\mathbf{T}_n$ to $\mathbf{T}^*_n$ by adding s_n to each open path of $\mathbf{T}_n$. Then extend $\mathbf{T}^*_n$ to $\mathbf{T}_{n+1}$ via the three steps of Lemma 7.4. As for each n, $\{s_i \mid i \leq n\}$ is fixed point satisfiable, the soundness theorem tells us that each tableaux $\mathbf{T}_i$ has at least one open path. Set $\mathbf{T} = \bigcup_{i \in \mathbb{N}} \mathbf{T}_i$. From the construction, it follows that each closed path of $\mathbf{T}$ is finite and that each open path is downwards saturated. Also note that $\mathbf{T}$ is a finitely branching tree with infinitely many points (signed sentences). According to *König's lemma* (see, for instance [50]), $\mathbf{T}$ has an infinite path $\mathcal{P}$. From the construction it follows that $\mathcal{P}$ is open, downwards saturated and that it contains all the elements of S. Hence, from Theorem 7.2 and Lemma 7.1 it follows that S is fixed point satisfiable.

We have now established completeness for arbitrary sets of sentences, as we will illustrate for the *st* case. Let Γ and Δ be arbitrary sets of sentences of L_T and suppose that $\Gamma \not\models^{st} \Delta$. This means that, for no *finite* $\Gamma' \subseteq \Gamma$ and $\Delta' \subseteq \Delta$ we have that $\Gamma' \vdash^{st} \Delta'$. From the finite completeness theorem, it follows that for every such Γ' and Δ' we have that $\{A^s_\sigma \mid \sigma \in \Gamma'\} \cup \{D^s_\sigma \mid \sigma \in \Delta'\}$ is fixed point satisfiable. From compactness, it follows that $\{A^s_\sigma \mid \sigma \in \Gamma\} \cup \{D^s_\sigma \mid \sigma \in \Delta\}$ is fixed point satisfiable. Hence $\Gamma \not\models^{st} \Delta$. Thus:

Theorem 7.5 With $i, j \in \{s, t\}$: $\vdash^{ij}$ is sound and complete w.r.t. $\models^{ij}$.
Proof: Given above. □

In the remainder of this section, we exploit the strict-tolerant calculus to make a couple of remarks on the classical behavior of $\models^{st}$. We show how an instructive proof of Fact 3 can be given by defining the *classical calculus* as a

sub-calculus of the strict-tolerant calculus. Before we do that, we define $\models^{cl}$, the classical consequence relation over ground models.

The relation $\models^{cl}$. Let, for any ground model M, $\mathbf{CL}(L_T, M)$ be the set of all classical valuations of L_T. A classical valuation V in $\mathbf{CL}(L_T, M)$ must respect the ground model M, i.e., $V(\sigma) = \mathcal{C}_M(\sigma)$ for all $\sigma \in L$, but it may give the truth predicate an arbitrary (classical) extension. Hence, a classical valuation may (and sometimes must, in the presence of Liar like statements) violate the identity of truth. $\mathbf{CL}(L_T)$ will denote the class of all classical valuations of L_T which respect some ground model M. That is:

$$V \in \mathbf{CL}(L_T) \Leftrightarrow V \in \mathbf{CL}(L_T, M) \text{ for some ground model } M$$

We now define $\models^{cl}$, "classical consequence in presence of ground models" by quantifying over all "ground model respecting classical interpretations of L_T", i.e, by quantifying over $\mathbf{CL}(L_T)$:

- $\Gamma \models^{cl} \Delta$ iff for every $V \in \mathbf{CL}(L_T)$:

$$\forall \alpha \in \Gamma : V(\alpha) = 1 \Rightarrow \exists \beta \in \Delta : V(\beta) = 1$$

The phrase "$\models^{st}$ is highly classical" may now be taken to be a paraphrase of: *if* $\Gamma \models^{cl} \Delta$ *then* $\Gamma \models^{st} \Delta$. In order to give an instructive proof of the latter claim, we define the *classical tableau calculus*.

The classical tableau calculus. We define the classical calculus as a (all too familiar) sub-calculus of the strict-tolerant calculus, exploiting only those rules of the strict-tolerant calculus that take us from strictly signed sentences to other strictly signed sentences and without the rules pertaining to the truth predicate. The closure conditions of the classical calculus are the same as those of the strict-tolerant calculus.

The relation $\vdash^{cl}$. We may use the classical tableau calculus to define the relation $\vdash^{cl}$ of "syntactic classical consequence in the presence of ground models" as follows: $\Gamma \vdash^{cl} \Delta$ iff: for some finite $\Gamma' \subseteq \Gamma, \Delta' \subseteq \Delta$ there is tableau starting with $\{A^s_\sigma \mid \sigma \in \Gamma'\} \cup \{D^s_\sigma \mid \sigma \in \Delta'\}$ that is closed according to the classical calculus.

In order to show that "$\models^{st}$ is highly classical" in a precise manner, we will exploit the following two lemma's:

Lemma 7.5 If $\Gamma \vdash^{cl} \Delta$ then $\Gamma \vdash^{st} \Delta$ (and hence $\Gamma \models^{st} \Delta$)
Proof: The claim that $\Gamma \vdash^{cl} \Delta$ then $\Gamma \vdash^{st} \Delta$ follows from the fact that the closure conditions of the classical rules are the same as those of the strict-tolerant calculus, that its rules are a subset of the rules of the strict-tolerant calculus and from the definition of $\vdash^{cl}$ and $\vdash^{st}$ in terms of the classical respectively strict-tolerant calculus. The addendum "and hence $\Gamma \models^{st} \Delta$" follows from soundness of $\vdash^{st}$ with respect to $\models^{st}$ □

Lemma 7.6 $\vdash^{cl}$ is sound and complete with respect to $\models^{cl}$
Proof: Follow a typical soundness and completeness proof for classical logic based on a 2-signed tableau calculus (see, e.g. [50]) and carry out some minor adjustments to account for ground models in a way similar to the proof of Theorem 7.2. □

We now have enough ammunition to prove the promised:

Proposition 7.1 If $\Gamma \models^{cl} \Delta$ then $\Gamma \models^{st} \Delta$
Proof: Let $\Gamma \models^{cl} \Delta$. According to Lemma 7.6, $\vdash^{cl}$ is complete with respect to $\models^{cl}$ and so we have that $\Gamma \vdash^{cl} \Delta$. From Lemma 7.5, it then follows that $\Gamma \models^{st} \Delta$. □

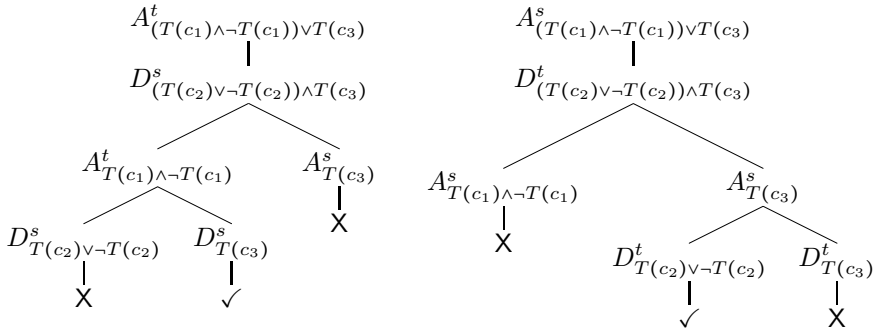
In the introduction, we illustrated some undesirable properties of $\models^{ss}$ and $\models^{tt}$; the first was shown to violate identity whereas the latter was shown to violate material modus ponens. In contrast, as $\models^{st}$ is highly classical (cf. Proposition 7.1) it satisfies both identity and material modus ponens. As the reader may have noticed, the relations $\models^{ss}$ and $\models^{tt}$ are complementary with respect to identity and material modus ponens: although $\models^{ss}$ violates identity, $\models^{tt}$ satisfies identity, and although $\models^{tt}$ violates material modus ponens, $\models^{ss}$ satisfies that property. These observations suggest that, more generally, the non-classical behavior of $\models^{ss}$ and $\models^{tt}$ may not be shared by their (set-theoretic) union $\models^{or}$:

- $\Gamma \models^{or} \Delta \Leftrightarrow \Gamma \models^{ss} \Delta$ or $\Gamma \models^{tt} \Delta$

The following example illustrates that this suggestion does not hold water: $\models^{or}$ is not highly classical in the sense of Proposition 7.1. First, observe that:

$$(T(c_1) \wedge \neg T(c_1)) \vee T(c_3) \models^{cl} (T(c_2) \vee \neg T(c_2)) \wedge T(c_3) \tag{7.19}$$

Proposition 7.1 implies that (7.19) also holds when $\models^{cl}$ is replaced by $\models^{st}$. However, (7.19) does not hold when $\models^{cl}$ is replaced by either $\models^{ss}$ or $\models^{tt}$, as the tableaux displayed below illustrate.



A check mark (✓) put below a branch indicates that (upon any further expansions of the tableau), the branch remains open, whereas a cross (X) indicates (upon further expansions of the tableau) the branch will close. Hence, the left tableau illustrates that $(T(c_1) \wedge \neg T(c_1)) \vee T(c_3)$ does not $\models^{ss}$ entail $(T(c_2) \vee \neg T(c_2)) \wedge T(c_3)$, whereas the right tableau illustrates that $(T(c_1) \wedge \neg T(c_1)) \vee T(c_3)$ does not $\models^{tt}$ entail $(T(c_2) \vee \neg T(c_2)) \wedge T(c_3)$.

7.4 Assertoric Semantics

The strict-tolerant calculus gives us a syntactic characterization of the four fixed point *consequence* relations. But, so one may ask, which sentences are, *actually* (strictly -tolerantly) assertible and or deniable? More precisely, given

a fixed ground model M, what is the assertoric status of the L_T sentences in strict-tolerant terms? In this section, we show that the strict-tolerant calculus suggests a very natural and precise answer to the actuality question. We give this answer in terms of *assertoric semantics*[6], which is a *semantic* version of the strict-tolerant calculus; assertoric semantics converts ground models into semantic valuations L_T. Let us explain the essential ideas of assertoric semantics and how it relates to the strict-tolerant calculus.

The closure conditions of the strict-tolerant calculus are thought of as the norms pertaining to (strict-tolerant) assertion and denial. Assertoric semantics acknowledges those norms and augments them with straightforward assertoric norms due to the ground model M, pertaining to sentences of L; if, with $\sigma \in Sen(L)$, $\mathcal{C}_M(\sigma) = 1$, then it is forbidden to (strictly or tolerantly) deny σ in M, while if $\mathcal{C}_M(\sigma) = 0$ then it is forbidden to (strictly or tolerantly) assert σ in M. Also, assertoric semantics slightly modifies the rules of the strict-tolerant calculus to acknowledge the fact that we are considering a fixed ground model $M = (D, I)$; the rules for the quantifiers now acknowledge the fact that the quantifiers range over D and the rules for the truth predicate are now applicable whenever an arbitrary constant $\overline{\sigma}$ *denotes* (as specified by I) the sentence σ. Further, the rules pertaining to identity statements are dropped (as their function is taken care of by M and the adjusted T-rules) while the rules for the propositional connectives remain the same. Below, we display the *assertoric rules of* L_T, i.e., the semantic counterpart of the rules of the strict-tolerant calculus as just discussed. An assertoric rule is either of *conjunctive type* $\sqcap$ or of *disjunctive type* $\sqcup$. Depending on its type, an assertoric rule is depicted in either one of the following two ways.

$$\frac{X^i_\sigma}{\Pi(X^i_\sigma)}\sqcup \qquad\qquad \frac{X^i_\sigma}{\Pi(X^i_\sigma)}\sqcap \tag{7.20}$$

The assertoric rules of L_T are, together with their type, displayed in the following table, where $i \in \{s, t\}$:

$\neg$	$\dfrac{A^i_{\neg\alpha}}{\{D^i_\alpha\}}\sqcap$	$\dfrac{D^i_{\neg\alpha}}{\{A^i_\alpha\}}\sqcup$
$\vee$	$\dfrac{A^i_{(\alpha\vee\beta)}}{\{A^i_\alpha, A^i_\beta\}}\sqcup$	$\dfrac{D^i_{(\alpha\vee\beta)}}{\{D^i_\alpha, D^i_\beta\}}\sqcap$
$\wedge$	$\dfrac{A^i_{(\alpha\wedge\beta)}}{\{A^i_\alpha, A^i_\beta\}}\sqcap$	$\dfrac{D^i_{(\alpha\wedge\beta)}}{\{D^i_\alpha, D^i_\beta\}}\sqcup$
$\exists$	$\dfrac{A^i_{\exists x\phi(x)}}{\{A^i_{\phi(x/c)} \mid c \in Con(L_T)\}}\sqcup$	$\dfrac{D^i_{\exists x\phi(x)}}{\{D^i_{\phi(x/c)} \mid c \in Con(L_T)\}}\sqcap$
$\forall$	$\dfrac{A^i_{\forall x\phi(x)}}{\{A^i_{\phi(x/c)} \mid c \in Con(L_T)\}}\sqcap$	$\dfrac{D^i_{\forall x\phi(x)}}{\{D^i_{\phi(x/c)} \mid c \in Con(L_T)\}}\sqcup$
T	$\dfrac{A^i_{T(\overline{\sigma})}}{\{A^i_\sigma\}}\sqcap$	$\dfrac{D^i_{T(\overline{\sigma})}}{\{D^i_\sigma\}}\sqcup$

[6] Assertoric semantics was first developed and used by Wintein [57]. The semantic valuation function used by [57] coincides with the *strict valuation function* as defined below.

Recall that, in the rules for the truth predicate, $\overline{\sigma}$ is an arbitrary constant which *denotes* the sentence σ. Hence, the assertoric rules are specified relative to a fixed ground model $M = (D, I)$. Further, observe that the assertoric rules for the quantifiers treat quantification substitutionally; for sake of simplicity, we assume that M is such that for every object d in D there is a constant c of L_T such that $I(c) = d$. The type ($\sqcup$ or $\sqcap$) distinction pertaining to assertoric sentences is used to define the notion of a *branch* of an assertoric sentence X^i_σ. A branch of X^i_σ is closely related to a completed path in a tableau which starts with X^i_σ. The set of all branches of X^i_σ is called the *assertoric tree* associated with X^i_σ, denoted $\mathfrak{T}^\sigma_{X^i}$. The assertoric tree of X^i_σ is closely related to the completed tableau which starts with X^i_σ.

Branches and Trees. A set B is a *branch* of X^i_σ just in case the following four conditions hold.

1. $X^i_\sigma \in B$
2. ($Y^i_\alpha \in B$ and Y^i_α has type $\sqcap$) $\Rightarrow$ $Z^i_\beta \in B$ for all $Z^i_\beta \in \Pi(Y^i_\alpha)$
3. ($Y^i_\alpha \in B$ and Y^i_α has type $\sqcup$) $\Rightarrow$ $Z^i_\beta \in B$ for some $Z^i_\beta \in \Pi(Y^i_\alpha)$
4. No strict subset of B satisfies conditions 1,2 and 3.

The assertoric tree $\mathfrak{T}^\sigma_{X^i}$ is the set of all branches of X^i_σ:

$$\mathfrak{T}^\sigma_{X^i} = \{B \mid \ B \text{ is a branch of } X^i_\sigma \ \}$$

As branches are sets of assertoric sentences, the closure conditions of the strict-tolerant calculus are applicable to them. As discussed above, assertoric semantics augments these closure conditions with closure conditions that are associated with the ground model M that is under consideration. Here is how.

Closure$_M$ conditions for branches and trees. A branch is said to be *ground$_M$ closed*, just in case:

$$\exists \sigma \in Sen(L) : \left(A^i_\alpha \in B \ \&\ \mathcal{C}_M(\sigma) = 0\right) \text{ or } \left(D^i_\alpha \in B \ \&\ \mathcal{C}_M(\sigma) = 1\right)$$

A branch is called *closed$_M$* just in case it is ground$_M$ closed or it is closed according to the closure conditions of the strict-tolerant calculus. An assertoric tree is called *closed$_M$* just in case all its branches are closed$_M$. Branches and trees that are not closed$_M$ are called *open$_M$*. Due to the structure of branches, the definition of the closure$_M$ conditions can be highly simplified. As an inspection of the assertoric rules reveals, branches have the property that either all their elements have sign X^s or all their elements have sign X^t. Hence, a branch will never be closed$_M$ due to the occurrence of A^s_σ (D^s_σ) and D^t_σ (A^t_σ). Further, the closure conditions of the strict-tolerant calculus pertaining to identity and A^t_σ-D^t_σ clashes are also taken care of by the notion of ground$_M$-closure. These observations allow for the following reformulation of the closure$_M$ conditions. With $B \in \mathfrak{T}^\sigma_{X^i}$, *$B$ is closed$_M$ just in case*:

$$B \text{ is ground}_M\text{-closed or } \exists \sigma \in Sen(L_T) : \{A^s_\sigma, D^s_\sigma\} \subseteq B \tag{7.21}$$

Hence, with B a branch of some tolerant assertoric tree $\mathfrak{T}^{\sigma}_{X^t}$: *$B$ is closed$_M$ just in case*:

$$B \text{ is ground}_M\text{-closed.} \tag{7.22}$$

Assertoric semantics thus associates four *assertoric trees* with each sentence σ: the strict assertion tree $\mathfrak{T}^{\sigma}_{A^s}$, the strict denial tree $\mathfrak{T}^{\sigma}_{D^s}$, and their tolerant counterparts $\mathfrak{T}^{\sigma}_{A^t}$ and $\mathfrak{T}^{\sigma}_{D^t}$. Intuitively, $\mathfrak{T}^{\sigma}_{A^s}$ keeps track of all the assertoric commitments that are associated with a strict assertion of σ in M, while the closure$_M$ of $\mathfrak{T}^{\sigma}_{A^s}$ indicates that it is not possible to live up to those commitments. When $\mathfrak{T}^{\sigma}_{A^s}$ is closed$_M$ we say that it is *forbidden* to strictly assert σ in M. When $\mathfrak{T}^{\sigma}_{A^s}$ is open$_M$, we say that it is *allowed* to strictly assert σ. The (closure$_M$ conditions pertaining to the) trees $\mathfrak{T}^{\sigma}_{D^s}$, $\mathfrak{T}^{\sigma}_{A^t}$ and $\mathfrak{T}^{\sigma}_{D^t}$ are interpreted similarly. The assertoric trees and the closure$_M$ conditions will be used to define the *strict* and the *tolerant* valuation of L_T in M.

The strict and tolerant valuation of L_T. The strict valuation $\mathcal{V}^s_M$ and the tolerant valuation $\mathcal{V}^t_M$ are defined in accordance with the following schema, where $i \in \{s,t\}$:

$$\mathcal{V}^i_M(\sigma) = \begin{cases} (1,0), & \mathfrak{T}^{\sigma}_{A^i} \text{ is open}_M \ \&\ \mathfrak{T}^{\sigma}_{D^i} \text{ is closed}_M \\ (1,1), & \mathfrak{T}^{\sigma}_{A^i} \text{ is open}_M \ \&\ \mathfrak{T}^{\sigma}_{D^i} \text{ is open}_M \\ (0,0), & \mathfrak{T}^{\sigma}_{A^i} \text{ is closed}_M \ \&\ \mathfrak{T}^{\sigma}_{D^i} \text{ is closed}_M \\ (0,1), & \mathfrak{T}^{\sigma}_{A^i} \text{ is closed}_M \ \&\ \mathfrak{T}^{\sigma}_{D^i} \text{ is open}_M \end{cases}$$

Let us give an example to illustrate our definitions. Let $M = (D, I)$ be a ground model such that $I(\lambda) = \neg T(\lambda)$ and $I(\tau) = T(\tau)$. With $i \in \{s,t\}$, we have that $\mathfrak{T}^{\neg T(\lambda)\vee T(\tau)}_{A^i} = \{A^i_1, A^i_2\}$ and $\mathfrak{T}^{\neg T(\lambda)\vee T(\tau)}_{D^i} = \{D^i\}$, where:

$$A^i_1 = \{A^i_{\neg T(\lambda)}, D^i_{T(\lambda)}, D^i_{\neg T(\lambda)}, A^i_{T(\lambda)}\} \qquad A^i_2 = \{A^i_{T(\tau)}\}$$

$$D^i = \{D^i_{\neg T(\lambda)}, A^i_{T(\lambda)}, A^i_{\neg T(\lambda)}, D^i_{T(\lambda)}, D^i_{T(\tau)}\}$$

As both A^s_2 and A^t_2 are open$_M$, so are $\mathfrak{T}^{\neg T(\lambda)\vee T(\tau)}_{A^s}$ and $\mathfrak{T}^{\neg T(\lambda)\vee T(\tau)}_{A^t}$. However, as D^s is closed$_M$ whereas D^t is open$_M$, it holds that $\mathfrak{T}^{\neg T(\lambda)\vee T(\tau)}_{A^s}$ is closed$_M$ while $\mathfrak{T}^{\neg T(\lambda)\vee T(\tau)}_{A^t}$ is open$_M$. Hence, we have that:

$$\mathcal{V}^s_M(\neg T(\lambda) \vee T(\tau)) = (1,0), \qquad \mathcal{V}^t_M(\neg T(\lambda) \vee T(\tau)) = (1,1)$$

It is left to the reader to establish that:

$$\mathcal{V}^s_M(\neg T(\lambda)) = (0,0), \quad \mathcal{V}^t_M(\neg T(\lambda)) = (1,1) \tag{7.23}$$

$$\mathcal{V}^s_M(T(\tau)) = (1,1), \quad \mathcal{V}^t_M(T(\tau)) = (1,1) \tag{7.24}$$

$$\mathcal{V}^s_M(\neg T(\tau)) = (1,1), \quad \mathcal{V}^t_M(\neg T(\tau)) = (1,1) \tag{7.25}$$

$$\mathcal{V}^s_M(T(\tau) \wedge \neg T(\tau)) = (0,1), \quad \mathcal{V}^t_M(T(\tau) \wedge \neg T(\tau)) = (1,1) \tag{7.26}$$

A couple of remarks concerning the interpretation of these valuations are in place. Equation (7.23) states that the Liar is neither strictly assertible nor strictly deniable, while it is both tolerantly assertible and deniable. Equation (7.24) states that the Truthteller is both strictly assertible and deniable and, also, that the Truetheller is both tolerantly assertible and deniable. Some caution

is in order here. According to STCT, there is nothing wrong with a sentence being both tolerantly assertible and deniable. In fact, the Liar is the canonical example of a sentence which is, according to STCT, both assertible and deniable. There may very well be objections against the very idea of the tolerant account of assertion and denial, but these need not concern us here. According to STCT, however, the strict assertion of a sentence rules out its strict denial. So, how should we interpret $\mathcal{V}^s_M(T(\tau)) = (1,1)$, which seems to tell us that the Truthteller is both strictly assertible and deniable? Well, as indicating that it is both allowed to strictly assert the Truthteller and also to strictly deny it, but that it is not allowed to do so "at the same time"; strictly asserting the Truthteller means giving up the ("a priori") possibility of strictly denying it, while strictly denying the Truthteller means giving up the ("a priori") possibility of strictly asserting it. According to assertoric semantics, strictly asserting and denying $T(\tau)$ "at the same time" is—given the assertoric rules for negation—tantamount to strictly asserting $T(\tau) \wedge \neg T(\tau)$. And equation (7.26) testifies that this is not allowed; $\mathcal{V}^s_M(T(\tau) \wedge \neg T(\tau)) = (0,1)$ indicates that it is not allowed to strictly assert $T(\tau) \wedge \neg T(\tau)$, while it is allowed to strictly deny it. So, where $\mathcal{V}^s_M$ valuates both $T(\tau)$ and $\neg T(\tau)$ as $(1,1)$ (cf. (7.24) (7.25)) it valuates their conjunction as $(0,1)$. This suggests that $\mathcal{V}^s_M$ is a non-compositional valuation function. In fact, $\mathcal{V}^s_M$ is a familiar 4-valued non-compositional valuation function, while $\mathcal{V}^t_M$ is a familiar compositional valuation function. As we will see below, $\mathcal{V}^s_M$ is equivalent to the function that Kripke [33] defined by quantifying over all (Strong Kleene) fixed points relative to a fixed ground model M, whereas $\mathcal{V}^t_M$ is equivalent to the (Strong Kleene) minimal fixed point over M.

The functions $\mathcal{K}^{\mathbf{4}}_M$ and $\mathcal{K}_M$. In [33], Kripke defines two valuation functions, which we will denote as $\mathcal{K}^{\mathbf{4}}_M$ and $\mathcal{K}_M$. The function $\mathcal{K}^{\mathbf{4}}_M$ is defined by quantifying over $\mathbf{FP}_M(L_T)$:

- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = (1,0) \Leftrightarrow \exists V_M : V_M(\sigma) = 1$ and $\nexists V_M : V_M(\sigma) = 0$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = (1,1) \Leftrightarrow \exists V_M : V_M(\sigma) = 1$ and $\exists V_M : V_M(\sigma) = 0$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = (0,0) \Leftrightarrow \nexists V_M : V_M(\sigma) = 1$ and $\nexists V_M : V_M(\sigma) = 0$
- $\mathcal{K}^{\mathbf{4}}_M(\sigma) = (0,1) \Leftrightarrow \nexists V_M : V_M(\sigma) = 1$ and $\exists V_M : V_M(\sigma) = 0$

The function $\mathcal{K}_M$ can be defined along the following lines. Define a partial order $\leq$ on $\mathbf{FP}_M(L_T)$ by stipulating that

$$V_M \leq V'_M \Leftrightarrow \forall \sigma \in Sen(L_T) : V_M(\sigma) = 1 \Rightarrow V'_M(\sigma) = 1$$

It can be shown that $(\mathbf{FP}_M(L_T), \leq)$ has a minimal element, V^{min}_M, which is called the *minimal fixed point valuation*. The function $\mathcal{K}_M$ translates V^{min}_M as having range $\{(1,0),(1,1),(0,1)\}$. That is:

$$\mathcal{K}_M(\sigma) = \begin{cases} (1,0), & V^{min}_M(\sigma) = 1; \\ (1,1), & V^{min}_M(\sigma) = \frac{1}{2}; \\ (0,1), & V^{min}_M(\sigma) = 0. \end{cases}$$

With the definition of $\mathcal{K}^{\mathbf{4}}_M$ and $\mathcal{K}_M$ in place, we can make the remark that $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$ are familiar functions precise. For, we have that:

Theorem 7.6 $\mathcal{V}^s_M = \mathcal{K}^{\mathbf{4}}_M$
Proof: See [63]. □

Theorem 7.7 $\mathcal{V}^t_M = \mathcal{K}_M$
Proof: See appendix. □

The *initial* assertoric possibilities with respect to L_T sentences in a ground model M are described by $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$. However, upon performing assertoric actions, we take up certain assertoric commitments, which (may) rule out certain other assertoric actions as forbidden. The latter process is described by the strict-tolerant calculus. For instance, with $T(\tau)$ the Truthteller, $\mathcal{V}^s_M(T(\tau)) = \mathcal{V}^t_M(T(\tau)) = (1,1)$ indicates that we can initially perform any of the four assertoric actions (A^s, D^s, A^t or D^t) with respect to $T(\tau)$. However, a strict assertion of the Truthteller rules out both a strong and a tolerant denial of it, whereas a tolerant denial of the Truthteller rules out a strict assertion of it. More generally, the transmission of assertoric possibilities due to (strict and tolerant) assertions and denials is captured by the strict-tolerant calculus. For, we have that:

$\Gamma \vdash^{st} \Delta \Leftrightarrow$ strictly asserting all of Γ rules out strictly denying all of Δ.
$\Gamma \vdash^{ss} \Delta \Leftrightarrow$ strictly asserting all of Γ rules out tolerantly denying all of Δ.
$\Gamma \vdash^{tt} \Delta \Leftrightarrow$ tolerantly asserting all of Γ rules out strictly denying all of Δ.
$\Gamma \vdash^{ts} \Delta \Leftrightarrow$ tolerantly asserting all of Γ rules out tolerantly denying all of Δ.

As $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$ describe the initial assertoric possibilities and as the strict-tolerant calculus describes the transmission of assertoric possibilities, it seems interesting to describe the dynamics of sequences of assertoric actions in a precise framework. Doing so, however, is beyond the scope of this paper.

7.5 Remarks on STCT

The plan of this section was announced in the introduction. We will mainly be considered with the following feature of STCT:

- STCT acknowledges four distinct (strict-tolerant) assertoric actions. The commitment to *four* distinct assertoric actions (rather than one) is, *prima facie*, an unattractive feature of STCT. However, Ripley [46] has argued that this need not be the case, for the strict-tolerant distinction is not a *primitive* one. Rather, the strict can be understood in terms of the tolerant or vice versa.

Looking at the strict-tolerant calculus, it is very natural (and we did so in the course of this paper) to interpret its four signs as *force indicators*. Doing so, however, we seem to be committed to the view that there are *four* primitive assertoric speech acts. Why then, does Ripley think of STCT as having bilateralist and not, say, *fourlateralist* commitments? Here is his answer:

> [The strict-tolerant distinction] it is not a primitive distinction; we can understand tolerant assertion and denial in terms of their strict cousins, as I've presented them here, or we can equally well understand strict in terms of tolerant. So long as we have a grip on one,

there is no difficulty in coming to understand the other.
(Ripley, [46, p20])

In a sense, this remark is to the point. But, so I will argue, not in the required sense. The remark is to the point *relative to a particular fixed point valuation* V_M. With V_M a fixed point valuation, we will say that σ *is strongly*$_{V_M}$ *assertible* just in case $V_M(\sigma) = 1$. The notion of a sentence being *strongly*$_{V_M}$ *deniable*, *tolerantly*$_{V_M}$ *assertible* and *tolerantly*$_{V_M}$ *deniable* are defined analogously. Relative to V_M, we can indeed understand strict in terms of tolerant (and vice versa), as:

$$\sigma \text{ is strongly}_{V_M} \text{ assertible} \Leftrightarrow \sigma \text{ is not tolerantly}_{V_M} \text{ deniable} \tag{7.27}$$

$$\sigma \text{ is strongly}_{V_M} \text{ deniable} \Leftrightarrow \sigma \text{ is not tolerantly}_{V_M} \text{ assertible} \tag{7.28}$$

Fair enough. But (7.27) and (7.28) only indicate that strict and tolerant can be understood in terms of one another if, given a ground model M, there would be a *privileged* fixed point valuation V_M^* which would inform us about the assertoric status of the L_T sentences. One way to argue that there must be such a privileged fixed point valuation is to give in to the intuition that Michael Kremer [29] calls *the supervenience of semantics*. Kremer contrasts the supervenience of semantics with another intuition, which he calls the *fixed point conception of truth*:

- *Supervenience of semantics*: Once all the empirical facts have been settled, so are all the semantic facts. In terms of our formal theory, the intuition becomes: for any given ground model, there is *exactly one* correct interpretation of the truth predicate. (Kremer [29, p238])

- *Fixed point conception of truth*: This criterion takes the notion of a fixed point to give the whole meaning of true. Or, in Kripke's words, the intuitive concept of truth is expressed by the formula: 'we are entitled to assert (or deny) of a sentence that it is true precisely under the circumstances when we can assert (or deny) the sentence itself'.
(Kremer [29, p241])

Clearly, due to its inferentialist commitments, STCT is committed to the fixed point conception of truth[7]: the meaning of true is given by the inferential (assertoric) rules of the truth predicate, and those do not single out a particular fixed point as privileged.

So, given STCT's inferential commitments, (7.27) and (7.28) do not suffice as a justification of the claim that the strict and the tolerant can be understood in terms of one another. In fact, the results of the previous section can be used to illustrate the insufficiency of (7.27) and (7.28) more concretely. Suppose that we are in an *initial situation*, as described at the end of the previous section. There, we saw that the strict and tolerant assertoric possibilities with respect

[7]If one thinks that, given a ground model M, there is exactly one correct fixed point valuation V_M^*, it would make sense to define a consequence relation by quantifying over the class of all V_M^* valuations. Thus, the fact that STCT defines its privileged consequence relation $\models^{st}$ by quantifying over *all* fixed point valuations testifies of its commitment to the fixed point conception of truth.

to L_T sentences in M are described by $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$ respectively. Clearly then, in an initial situation the strict-tolerant conversion cannot proceed in line with (7.27) and (7.28); $\mathcal{V}^t_M$ and $\mathcal{V}^s_M$ are distinct functions and $\mathcal{V}^s_M$ is not even a fixed point valuation. Thus, an initial situation suggests that ST is committed to taking the strict-tolerant distinction as a primitive one and hence, to acknowledge that there are four primitive assertoric speech acts.

But not so fast. For, so one may argue, the above objections notwithstanding, ST is not forced to give up its claim that we can get away with two primitive assertoric speech acts. Here is an argument which purports to establish just that.

Argument 1 There are only two primitive assertoric speech acts: assertion and denial. However, there are two distinct *norms* that govern the practice of asserting and denying sentences: a strict and a tolerant one. For instance, the distinction between a strict and tolerant assertion is not a force distinction between two speech acts, but rather a distinction in the norm according to which a single speech act (assertion) is qualified as allowed / forbidden. In fact, this is clearly illustrated by the definition of $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$, as these functions are (modulo an insignificant difference in signs) obtained from the same assertoric trees. The only difference is that $\mathcal{V}^s_M$ is obtained by judging the assertoric trees to be open / closed in terms of the strict norm of assertion, whereas the closure judgements that induce $\mathcal{V}^t_M$ arise from the tolerant norm.

Although this argument seems plausible when we restrict our attention to $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$, the closure conditions of the strict-tolerant calculus reveal that it is incorrect. For, recall that a sufficient condition for the closure of a set of assertoric sentences S is that:

$$\{A^s_\sigma, D^t_\sigma\} \subseteq S \text{ or } \{A^t_\sigma, D^s_\sigma\} \subseteq S \tag{7.29}$$

Hence, strictly (tolerantly) asserting σ rules out tolerantly (strictly) denying σ and vice versa. In the definition of $\mathcal{V}^s_M$ and $\mathcal{V}^t_M$, (7.29) played no role as these functions do not consider the interactions between strict and tolerant. However, (7.29) is a condition *sine qua non* for our syntactic characterization of the fixed point consequence relations via the strict-tolerant calculus. Thus, according to the strict-tolerant calculus, there are no two separate norms—a tolerant and a strict one—but rather there is a single norm that governs the practice of strictly and tolerantly asserting and denying sentences. According to the strict-tolerant calculus, there are four distinct primitive assertoric speech acts which are governed by a single norm that is formally represented by its closure conditions. Hence, argument 1 fails.

Here is another argument that a ST proponent may invoke for this claim that the strict and tolerant are on a par.

Argument 2 What is really at the heart of ST is the commitment to (a syntactic characterization of) $\vDash^{st}$. The strict and tolerant are on a par as we can characterize this relation in two ways: by putting constraints on strict assertion and denial (only) or by putting constraints on tolerant assertion and denial (only).

According to the strict-tolerant calculus, this claim is plainly wrong. Clearly, we can characterize $\models^{st}$ by putting constraints on strict assertion and denial only, as the definition of $\vdash^{st}$, which can be paraphrased as **A**, testifies.

A Strictly asserting all premisses rules out strictly denying all consequences.

However, we can't do so by putting constraints on tolerant assertion and denial only. According to the strict-tolerant calculus, the strict and tolerant are, with respect to $\models^{st}$, not on a par. Rather, strict assertions and denials have a privileged status. It may be argued that the discrepancy between the strict and tolerant alluded to is a phenomenon that is specific to the strict-tolerant calculus. Thus, there may be other calculi that allow us to characterize $\models^{st}$ and according to which argument 2 can be sustained. As of yet, I have no argument which rules out such a possibility in principle. Then again, I take it that the strict-tolerant calculus is a very *natural* way of characterizing the (consequence relations induced by) strict and tolerant assertions and denials. The naturalness of the strict-tolerant calculus at least *suggests* that argument 2 is false. At any rate, it shifts the burden of proof to someone who claims that "the strict and tolerant are on a par".

In what follows, we simply assume that we have established that according to STCT, the strict and tolerant cannot be conceived of as being on a par. In other words, we take it that STCT is committed to acknowledge that there are four primitive assertoric speech acts. We now ask the question whether, in light of this commitment, STCT's self-declared bilateralism has to be reconsidered. Inferentialism is the view that meanings are to be explained in terms of which inferences are valid, while *bilateralism* is a species of interentialism according to which the validity of inferences is to explained in terms of conditions on assertion and denial. According to STCT, the relation $\models^{st}$ specifies which inferences are valid. As we pointed out, $\models^{st}$ validity can be explained in terms of conditions on (strict) assertion and (strict) denial. Hence, in this sense, STCT can clearly uphold its commitment to bilateralism. On the other hand, it is not outrageous to assert that an advocate of *bi*-lateralism is committed to view that there are only two primitive assertoric speech acts. On this understanding of bilateralism, STCT has, per definition, to give up its commitment to bilateralism. Thus, whether or not STCT has to give up its commitment to bilateralism depends on what you mean by bilateralism. At any rate, STCT has to give up its claim that the strict tolerant distinction is not a primitive one. Or so we argued. The philosophical implications of a commitment to four distinct assertoric speech acts are beyond the scope of this paper.

7.6 Syntactic approaches to Strong Kleene Truth

In this section, we will compare our characterization of the fixed point consequence relations with other approaches that characterize (some of those) relations.

In [29], Michael Kremer gives a 2-sided sequent calculus which is sound and complete with respect to the derived fixed point consequence relation $\models^{\&}$, which he defines by stipulating that $\Gamma \models^{\&} \Delta$ iff both $\Gamma \models^{ss} \Delta$ and $\Gamma \models^{tt} \Delta$. Although

Kremer points out that his results can be modified to deal with $\models^{ss}$ and $\models^{tt}$ as well, he does not discuss $\models^{st}$ and $\models^{ts}$. In fact, Kremer is not concerned with the strict-tolerant distinction at all. In comparison with Kremer then, our approach is more general, as it captures $\models^{st}$ and $\models^{ts}$ as well. Also, our approach is motivated by distinct philosophical considerations, the strict-tolerant distinction, which are reflected directly in our calculus via our four signs.

Ripley [45] gives a unified 3-sided sequent calculus which can be used to define syntactic consequence relations that capture (i.e., are sound and complete with respect to) $\models^{st}$, $\models^{ss}$, $\models^{tt}$ and $\models^{ts}$. As such, it is natural to compare our approach to that of [45]. Here are, what I consider, the two most significant distinctions between our approach and that of [45].

1. The three positions in a 3-sided sequent of the calculus of [45] correspond to the three distinct semantic values, 1, $\frac{1}{2}$ and 0, of a fixed point valuation. Our four signs, in contrast, correspond directly to the four strict-tolerant assertoric actions that are constitutive for the strict-tolerant interpretation of fixed point valuations. In that sense, our calculus is better suited (in fact, tailor made) to study the notions of strict / tolerant assertion and denial than the sequent calculus of [45]. Further, an important benefit of the strict-tolerant calculus is its connection with assertoric semantics, as explained in Section 3. As a consequence, the strict-tolerant calculus has the possibility to assess STCT's claim in a precise manner, as we illustrated in Section 4.

2. Strictly speaking, [45] is not concerned with the four fixed point consequence relations as we defined them. The reason is that for us, a fixed point valuation is defined as a valuation that is *i* Strong Kleene, *ii* respects a *classical* ground model M and *iii* respects the identity of truth. In contrast, the four consequence relations considered by [45] are defined by quantifying over all Strong Kleene valuations of L_T which respect the identity of truth; condition *ii* is dropped. Clearly, dropping *ii* radically changes the extensions of the consequence relations. To see this, observe that for us a sentence of form $P(c) \vee \neg P(c)$ (where $P \in L$) is always strictly assertible, whereas this is not the case according to [45]. Now, it is common practice to define theories of truth relative to a classical ground model. More importantly, it seems also methodologically preferable to do so: the reason that a truth-free sentence valuates as $\frac{1}{2}$ cannot involve the behavior of the truth predicate. As such, a study of this behavior better abstracts away from these reasons. Accordingly, we feel that it is worthwhile to study the four consequence relations with classical ground models around. In this sense, our approach resembles that of Kremer [29], who also requires that a fixed point valuation respects a classical ground model.

A last point of comparison is another paper of Ripley. In [46], Ripley articulates STCT and, in order to do so, he restricts his attention to $\models^{st}$. To characterize $\models^{st}$ syntactically, Ripley does not rely on the 3-sided sequent calculus of [45], but rather, he presents a 2-sided sequent calculus which is sound and complete with respect to $\models^{st}$. The 2-sided calculus can, in the usual sense, be considered as the sequent calculus variant of $\vdash^{st}$ as defined in this paper: signing a sentence with A^s (D^s) in our calculus corresponds with placing that sentence on the left (right) in the calculus of [46]. The characterization of $\models^{st}$ as in [46] takes away the qualms (mentioned in 1 above) we have with respect to the 3-sided sequent approach of [45]. However, as the results in this paper point out, the natural approach to $\models^{st}$ by [46] does not carry over to $\models^{ss}$ and $\models^{tt}$ (or $\models^{ts}$) . The reason for this is that the *calculus* of [46] only recognizes a

single assertoric sense whereas, to characterize all four fixed point relations in (strict-tolerant) assertoric terms, a calculus needs to distinguish between strict and tolerant assertion and denial. Further, [46] is, just like [45], not considered with fixed point valuations that are induced by ground models.

7.7 Concluding remarks

We presented a tableau system, *the strict-tolerant calculus*, and used it to characterize all four fixed point consequence relations over L_T ($\models^{st}$, $\models^{ss}$, $\models^{tt}$ and $\models^{ts}$) in a uniform manner. Next, we showed how a semantic version of the strict-tolerant calculus, called *assertoric semantics*, can be used to capture the (strict-tolerant) assertoric status of the L_T sentences relative to a particular ground model. By exploiting the strict-tolerant calculus and assertoric semantics, we then argued that the strict-tolerant distinction is, pace Ripley, a primitive distinction. We concluded by indicating some important (methodological) distinctions between the strict-tolerant calculus and other calculi that are sound and complete with respect to (some of the) fixed point consequence relations.

Appendix

In this appendix, we will prove Theorem 7.7 of Section 3. Our proof will exploit a proof of [63], which showed how to represent the (Strong Kleene) minimal fixed point in terms of the function $\mathcal{V}^{gr}_M$, defined in terms of the *method of closure games*. We will show that $\mathcal{V}^{t}_M$ is equivalent to $\mathcal{V}^{gr}_M$, from which it follows—by the results of [63]—that $\mathcal{V}^{t}_M$ is equivalent to the minimal fixed point. We will first present the definition of $\mathcal{V}^{gr}_M$ via the method of closure games, which relies heavily on the assertoric rules for L_T as presented in Section 3. Here are the essential concepts involved in the definition of $\mathcal{V}^{gr}_M$.

Strategies, (grounded) expansions, closure conditions, valuations
1. A *strategy for player* $\sqcup$ is a function f which maps each X^t_σ of[8] type $\sqcup$ to one element of $\Pi(X^t_\sigma)$. The set of all strategies of player $\sqcup$ is denoted by $\mathcal{F}$.
2. A *strategy for player* $\sqcap$ is a function g which maps each X^t_σ of type $\sqcap$ to one element of $\Pi(X^t_\sigma)$. The set of all strategies of player $\sqcap$ is denoted by $\mathcal{G}$.
3. With $f \in \mathcal{F}$, $g \in \mathcal{G}$ and X^t_σ a tolerant assertoric sentence, $\mathsf{exp}(X^t_\sigma, f, g)$ denotes *the expansion of* X^t_σ *by* f *and* g. $\mathsf{exp}(X^t_\sigma, f, g)$ is an infinite—whenever we hit an atomic sentence that is not a sentential truth ascription we keep on repeating it—sequence of AD sentences whose first element is X^t_σ and whose successor relation respects f and g. As an example, with $g \in \mathcal{G}$ such that $g(A^t_{P(c_1)\wedge P(c_2)}) = A^t_{P(c_2)}$, $\mathsf{exp}(A^t_{P(c_1)\wedge P(c_2)}, f, g)$ is:

$$A^t_{P(c_1)\wedge P(c_2)}, A^t_{P(c_2)}, A^t_{P(c_2)}, A^t_{P(c_2)}, \ldots$$

The set of all expansions in M is[9] denoted by EXP_M.
4. A *closure condition* $\dagger(M) = \{O^\dagger_M, C^\dagger_M\}$ is a bipartition of EXP_M into the sets

[8] The method of closure games does not rely on the strict-tolerant distinction. Here, we present the method in terms of tolerant assertoric sentences (only) to make the connection with $\mathcal{V}^t_M$ more direct.

[9] The assertoric rules for truth testify that the set of all expansions depends on the ground model under consideration.

$O^{\dagger}_M \neq \varnothing$, consisting of the *open*$_{\dagger}$ *expansions in* M, and $C^{\dagger}_M \neq \varnothing$, containing the *closed*$_{\dagger}$ *expansions* in M.

5. A closure condition $\dagger(M) = \{O^{\dagger}_M, C^{\dagger}_M\}$ gives rise to closure conditions for AD sentences:

$$O^{\dagger}_M(X^t_\sigma) \Leftrightarrow \exists f \in \mathcal{F} \forall g \in \mathcal{G} : \mathsf{exp}(X^t_\sigma, f, g) \in O^{\dagger}_M$$

$$C^{\dagger}_M(X^t_\sigma) \Leftrightarrow \text{ not } O^{\dagger}_M(X^t_\sigma)$$

6. The closure conditions for AD sentences are used to induce $\mathcal{V}^{\dagger}_M$:

$$\mathcal{V}^{\dagger}_M(\sigma) = \begin{cases} (1,0), & O^{\dagger}_M(A^t_\sigma) \text{ and } C^{\dagger}_M(D^t_\sigma) \\ (1,1), & O^{\dagger}_M(A^t_\sigma) \text{ and } O^{\dagger}_M(D^t_\sigma) \\ (0,0), & C^{\dagger}_M(A^t_\sigma) \text{ and } C^{\dagger}_M(D^t_\sigma) \\ (0,1), & C^{\dagger}_M(A^t_\sigma) \text{ and } O^{\dagger}_M(D^t_\sigma) \end{cases} \tag{7.30}$$

7. An expansion is *grounded* just in case it contains, for some $\alpha \in At(L)$, X^t_α; we say that X^t_α is the *ground* of the expansion. Expansions that are not grounded are called *ungrounded*. Grounded expansions are either *correct in* M or *incorrect in* M. A (grounded) expansion is *correct in* M just in case its ground X^t_α is such that:

$$X = A \ \& \ \mathcal{C}_M(\alpha) = 1 \text{ or } X = D \ \& \ \mathcal{C}_M(\alpha) = 0$$

$G^{cor}_M \subseteq \mathsf{EXP}_M$ is the set consisting of all expansions which are grounded and correct in M. $G^{inc}_M \subseteq \mathsf{EXP}_M$ is the set consisting of all expansions which are grounded and incorrect in M.

The function $\mathcal{V}^{gr}_M$ is induced, in accordance with (7.30), by stipulating that all and only the expansions of G^{cor}_M are open, i.e., $O^{gr}_M = G^{cor}_M$. In [63], it was shown that $\mathcal{V}^{gr}_M : Sen(L_T) \to \{(1,0),(0,0),(0,1)\}$ is equivalent to the minimal fixed point $V^{min}_M : Sen(L_T) \to \{1, \frac{1}{2}, 0\}$ by taking $(1,0)$, $(0,0)$ and $(0,1)$ to be abbreviations of 1, $\frac{1}{2}$ and 0 respectively.

Here, we define the dual of $\mathcal{V}^{gr}_M$, denoted $\mathcal{V}^{\overline{gr}}_M$, by stipulating that all and only the expansions of G^{inc}_M are closed, i.e., $C^{\overline{gr}}_M = G^{inc}_M$. Due to the fact that a branch (as defined in Section 3) is set of expansions, we have that:

Lemma 7.7 $\mathcal{V}^{\overline{gr}}_M = \mathcal{V}^{t}_M$

Let $\mathfrak{T}^{\sigma}_{X^t}$ be a tolerant assertoric tree of σ. Clearly, it suffices to show that:

$$\mathfrak{T}^{\sigma}_{X^t} \text{ is open } \Leftrightarrow \exists f \in \mathcal{F} \forall g \in \mathcal{G} : \mathsf{exp}(X^t_\sigma, f, g) \in O^{\overline{gr}}_M,$$

which follows from a comparison of the closure conditions for branches with $C^{\overline{gr}}_M$ and by observing that a branch of $\mathfrak{T}^{\sigma}_{X^t}$ can be obtained by fixing an $f' \in \mathcal{F}$ and by putting all assertoric sentences that occur on some expansion that is contained in $\{\mathsf{exp}(X^t_\sigma, f', g) \mid g \in \mathcal{G}\}$ into a single set. □

We will show that, modulo a "$(0,0)$ – $(1,1)$ conversion", $\mathcal{V}^{gr}_M : Sen(L_T) \to \{(1,0),(0,0),(0,1)\}$ and $\mathcal{V}^{\overline{gr}}_M : Sen(L_T) \to \{(1,0),(1,1),(0,1)\}$ are identical. As $\mathcal{V}^{gr}_M$ is known to be the minimal fixed point, this proves Theorem 7.7.

Lemma 7.8 $\mathcal{V}_M^{gr}$ **and** $\mathcal{V}_M^{\overline{gr}}$ **are equivalent modulo a "**$(0,0)-(1,1)$ **conversion".**
Proof: First, note that it suffices to show that for any $\sigma \in Sen(L_T)$:

i $\mathcal{V}_M^{gr}(\sigma) = (1,0) \Leftrightarrow \mathcal{V}_M^{\overline{gr}}(\sigma) = (1,0)$

ii $\mathcal{V}_M^{gr}(\sigma) = (0,1) \Leftrightarrow \mathcal{V}_M^{\overline{gr}}(\sigma) = (0,1)$

With X_σ^t a (tolerant) assertoric sentence, $\hat{X}_\sigma^t$ will denote its *AD inverse*: $\hat{A}_\sigma^t = D_\sigma^t$ and $\hat{D}_\sigma^t = A_\sigma^t$. With S a set of (tolerant) assertoric sentences. $\hat{S}$ will denote the *AD inverse* of S, i.e. the set consisting of the *AD* inverses of the elements of S. As shown in [63] in more detail, symmetry considerations reveal that:

$$\exists f \forall g \; \mathsf{exp}(X_\sigma^t, f, g) \in S \Leftrightarrow \exists g \forall f \; \mathsf{exp}(\hat{X}_\sigma^t, f, g) \in \hat{S} \tag{7.31}$$

$i \Rightarrow$ Suppose that $\mathcal{V}_M^{gr}(\sigma) = (1,0)$, i.e., that $O_M^{gr}(A_\sigma^t)$ and $C_M^{gr}(D_\sigma^t)$. As $O_M^{gr} \subseteq O_M^{\overline{gr}}$, $O_M^{gr}(A_\sigma^t)$ implies that $O_M^{\overline{gr}}(A_\sigma^t)$. Further, $O_M^{gr}(A_\sigma^t)$ means that player $\sqcup$ has a strategy which ensures that the expansion of A_σ^t will end up in G_M^{cor}. As the *AD* inverse of G_M^{cor} is G_M^{inc}, this implies, via (7.31), that player $\sqcap$ has a strategy g which ensures that the expansion of D_σ^t will end up in G_M^{inc}. But this means that player $\sqcup$ does *not* have a strategy which ensures that the expansion of D_σ^t will end up in $\mathsf{EXP}_M - G_M^{inc} = O_M^{\overline{gr}}$. Thus, $C_M^{\overline{gr}}(D_\sigma^t)$ and so $\mathcal{V}_M^{\overline{gr}}(\sigma) = (1,0)$.
$i \Leftarrow$ Suppose that $\mathcal{V}_M^{\overline{gr}}(\sigma) = (1,0)$, i.e., that $O_M^{\overline{gr}}(A_\sigma^t)$ and $C_M^{\overline{gr}}(D_\sigma^t)$. As $O_M^{gr} \subseteq O_M^{\overline{gr}}$, $C_M^{\overline{gr}}(D_\sigma^t)$ implies that $C_M^{gr}(D_\sigma^t)$. Further, from $C_M^{\overline{gr}}(D_\sigma^t)$, it follows that $\sqcap$ has a strategy g which ensures that the expansion of D_σ^t will end up in G_M^{inc}. This implies, via (7.31) and as G_M^{cor} is the *AD* inverse of G_M^{inc}, that player $\sqcup$ has a strategy f which ensures that the expansion of D_σ^t will end up in G_M^{cor}. Hence $O_M^{gr}(A_\sigma^t)$ and so $\mathcal{V}_M^{gr}(\sigma) = (1,0)$.
ii Just like *i*. □

Theorem 7.7 $\mathcal{V}_M^t = \mathcal{K}_M$
Proof: From Lemma 7.7 and Lemma 7.8. □

Chapter 8

A Calculus for Belnap's Logic in Which Each Proof Consists of Two Trees[1]

8.1 Abstract

In this paper we introduce a Gentzen calculus for (a functionally complete variant of) Belnap's logic in which establishing the provability of a sequent in general requires *two* proof trees, one establishing that whenever all premises are true some conclusion is true and one that guarantees the falsity of at least one premise if all conclusions are false. The calculus can also be put to use in proving that one statement *necessarily approximates* another, where necessary approximation is a natural dual of entailment. The calculus, and its tableau variant, not only capture the classical connectives, but also the 'information' connectives of four-valued Belnap logics. This answers a question by Avron.

8.2 Introduction

In logics based on Belnap's [6, 7] well-known bilattice *FOUR* (see Figure 8.2) everything gets doubled. The two truth values of classical logic are replaced by their four possible combinations (we write **t** for 'true and not false', **f** for 'false and not true', **n** for 'neither true nor false', and **b** for 'both true and false'),[2] and there are two natural orderings on these values instead of the single classical ordering on {true, false}. One of these orderings, $\leq_t$, depicted in the Hasse diagram for the logical lattice L4 in Figure 8.2, is connected to the *degree of truth* a statement may assume; the other, $\leq_k$, the ordering in the approximation lattice A4, to its *degree of definedness*.[3] Four values bring more

[1] This paper is joint work with Reinhard Muskens.

[2] Wintein [57] gives an alternative reading of Belnap's four values in terms of the *assertibility* and *deniability* of statements.

[3] Ginzberg [20] considers a general theory of bilattices, but we will stick to the logic based on Belnap's *FOUR* here. For general information about bilattices, see Fitting's papers, e.g. [17].

truth functions with them than two do and this leads to a doubling of logical operators.[4] The classical $\neg$, $\wedge$ and $\vee$ are now naturally complemented with duals $-$ ('conflation'), $\otimes$ ('consensus') and $\oplus$ ('gullibility'), and in a predicate logical setting the quantifiers $\forall$ and $\exists$ moreover come with cousins Π and Σ.[5] Also, while in classical logic one can define entailment either as transmission of truth or, completely equivalently, transmission of non-falsity ('if the conclusion is false one of the premises must be'), these two notions come apart in the four-valued setting, since there is transmission of truth but not of non-falsity from $\mathbf{t}$ to $\mathbf{b}$, for example. Entailment is naturally defined by stipulating that $\varphi \vDash \psi$ if and only if the values of φ and ψ are in the $\leq_t$ ordering in every model (for every assignment) and this boils down to requiring that both forms of transmission must hold from φ to ψ.[6]

The doublings do not stop here. Entailment itself also obtains a natural dual, for, replacing $\leq_t$ in the above definition by $\leq_k$, we can say that φ *necessarily approximates* ψ, $\varphi \bowtie \psi$, if and only if the values of φ and ψ are in the $\leq_k$ ordering in every model. We feel that this notion of necessary approximation carries some interest given the pivotal role of the approximation (or 'knowledge') ordering in the semantics of programming languages.

The main purpose of this paper is a simple one. We want to add one more doubling to the ones mentioned already by giving a proof system for four-valued predicate logic in which each proof consists of *two* Gentzen proof trees, one establishing transmission of truth, the other transmission of non-falsity. The system can also be used to show that necessary approximation holds. In that case one proof tree again corresponds to transmission of truth but the other to transmission of falsity, not non-falsity. While Muskens [38] presents a Gentzen calculus in which only one proof tree is needed to establish provability, and while one tree may be thought to be nicer than two, this advantage is offset by the fact that the system of [38] is obviously biased towards the L4 ordering, as

[4] In Belnap [6, 7] only the classical operators are considered.

[5] See Fitting [17] for further motivation of these operators. One possible application of $\otimes$ and $\oplus$ that Fitting mentions is that they could be part of a logic programming language designed for distributed implementation, a suggestion that is quite in line with Belnap's original motivation.

[6] This is the notion of entailment considered in Belnap [6, 7], but not that of Arieli & Avron [1], who use a single-barrelled notion. The two notions of entailment are co-extensional on sets of formulas based on classical connectives only, but not on formulas based on a functionally complete set of connectives or on a set of connectives that expresses all $\leq_k$-monotone functions. Belnap [6, p43] is quite clear about his views on the connection between entailment and the lattice L4. Considering the question when an argument in his logic is a good one, he writes:

> The abstract answer relies on the *logical* lattice we took so much time to develop. It is: entailment goes uphill. That is, given any sentence A and B (compounded from variables by negation, conjunction and disjunction), we will say that A *entails* or *implies* B just in case for each assignment of one of the four values to the variables, the value of A does not exceed (is less-than-or-equal-to) the value of B.

On the same page Belnap refers to Dunn [13], who shows that preservation of truth and preservation of non-falsity coincide for classical sentences, but he nevertheless insists on defining entailment as preservation of truth *and* non-falsity:

> But I agree with the spirit of a remark of Dunn's, which suggests that the False really is on all fours with the True, so that it is profoundly natural to state our account of "valid" or "acceptable" inference in a way which is neutral with respect to the two.

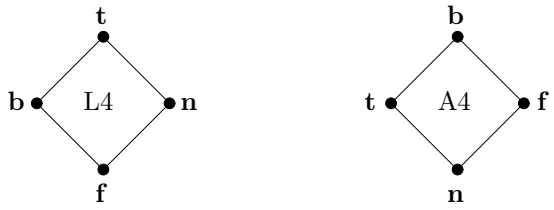


Figure 8.1: Belnap's bilattice *FOUR* depicted in terms of its constituting lattices L4 and A4.

its sequent rules for $\wedge$, $\vee$, $\forall$ and $\exists$ are natural and familiar, while those for $\otimes$, $\oplus$, Π and Σ come out rather convoluted. The present calculus is more natural, with sequent rules in the second group completely dual to those in the first. It is also more natural in the sense that the 'structural elements' of [38] can be done away with (without giving up analiticity).

Naturalness is also our defense against the objection that there are now uniform methods by which signed (analytic) proof systems for finite valued logics can be obtained (see e.g. Baaz et al. [4]). While this is clearly an important general result, the proof systems that are obtained in particular cases are sometimes unnecessarily complicated and the system generated for Belnap's logic using the method of [4] is a case in point. The binary connectives, for instance, are provided with tableau rules that have up to four disjunctive clauses, while these clauses themselves may consist of sets of signed statements rather than of single signed statements. Consider, for example, the tableau rules for $\mathbf{b} : \varphi \wedge \psi$ and $\mathbf{f} : \varphi \wedge \psi$ obtained in this manner.

$$\frac{\mathbf{b} : \varphi \wedge \psi}{\mathbf{t} : \varphi, \mathbf{b} : \psi \mid \mathbf{b} : \varphi, \mathbf{t} : \psi \mid \mathbf{b} : \varphi, \mathbf{b} : \psi} \qquad \frac{\mathbf{f} : \varphi \wedge \psi}{\mathbf{f} : \varphi \mid \mathbf{f} : \psi \mid \mathbf{n} : \varphi, \mathbf{b} : \psi}$$

The proof system of the present paper exploits the fact that **t**, **f**, **n** and **b** are best thought of as combinations of truth values. We choose our four signs to capture the "underlying" values of (non-)truth and (non-)falsity and, in doing so we arrive at a proof system that is tailor made for Belnap's logic. In the tableau variant of our system, a signed tableau rule for a binary connective is either of disjunctive or conjunctive type and always involves exactly two immediate descendants. In this sense, our system resembles the tableau calculus for first order logic of Smullyan [50].

The remainder of the paper is organized as follows. The next section gives the syntax and semantics of Belnap's logic; Section 8.4 introduces the 'two trees' proof system; Section 8.5 answers a question by Avron by discussing a tableau variant of this proof system which extends that of Avron [3]; and Section 8.6 is a conclusion. An appendix gives a series of Gentzen rules for defined operators.

8.3 $\mathbf{L_4}$: Syntax and Semantics

In setting up the four-valued predicate logic $\mathbf{L_4}$ we will by and large follow Muskens [38], and refer to this paper for discussion of the concepts involved.

The set of *formulas* of $\mathbf{L_4}$ is defined just as it is done in standard predicate logic, except that $-$ and $\otimes$ are added to the familiar $\neg$, $\wedge$, $\forall$ and $=$. A *model* is a pair $\langle \mathcal{D}, \mathcal{I} \rangle$ where $\mathcal{D} \neq \varnothing$ and $\mathcal{I}$ is a function with as domain the language (set of non-logical constants) $\mathcal{L}$, such that $\mathcal{I}(f)$ is an n-ary function on $\mathcal{D}$ if $f \in \mathcal{L}$ is an n-ary function symbol and $\mathcal{I}(R)$ is a *pair* of n-ary relations on $\mathcal{D}$ if $R \in \mathcal{L}$ is an n-ary relation symbol. We denote the first element of this pair as $\mathcal{I}^+(R)$, the second element as $\mathcal{I}^-(R)$. We use a to denote a (variable) assignment and a^x_d to denote the assignment that is like a except for assigning d to x. The value of a term t in a model $\mathcal{M}$ under an assignment a is defined in the usual way and written as $[\![t]\!]^{\mathcal{M},a}$, or $[\![t]\!]^{\mathcal{M}}$ if t is closed.

Definition 8.1 *We define the three-place relations $\mathcal{M} \vDash \varphi[a]$ (formula φ is* true *in model $\mathcal{M}$ under assignment a) and $\mathcal{M} \nvDash \varphi[a]$ (φ is* false *in $\mathcal{M}$ under a) as follows.*

1. $\mathcal{M} \vDash Rt_1 \dots t_n[a] \Leftrightarrow \langle [\![t_1]\!]^{\mathcal{M},a}, \dots, [\![t_n]\!]^{\mathcal{M},a} \rangle \in \mathcal{I}^+(R)$,
   $\mathcal{M} \nvDash Rt_1 \dots t_n[a] \Leftrightarrow \langle [\![t_1]\!]^{\mathcal{M},a}, \dots, [\![t_n]\!]^{\mathcal{M},a} \rangle \in \mathcal{I}^-(R)$;
2. $\mathcal{M} \vDash t_1 = t_2[a] \Leftrightarrow [\![t_1]\!]^{\mathcal{M},a} = [\![t_2]\!]^{\mathcal{M},a}$,
   $\mathcal{M} \nvDash t_1 = t_2[a] \Leftrightarrow [\![t_1]\!]^{\mathcal{M},a} \neq [\![t_2]\!]^{\mathcal{M},a}$;
3. $\mathcal{M} \vDash \neg\varphi[a] \Leftrightarrow \mathcal{M} \nvDash \varphi[a]$,
   $\mathcal{M} \nvDash \neg\varphi[a] \Leftrightarrow \mathcal{M} \vDash \varphi[a]$;
4. $\mathcal{M} \vDash -\varphi[a] \Leftrightarrow \mathcal{M} \nvDash \varphi[a]$,
   $\mathcal{M} \nvDash -\varphi[a] \Leftrightarrow \mathcal{M} \nvDash \varphi[a]$;
5. $\mathcal{M} \vDash \varphi \wedge \psi[a] \Leftrightarrow \mathcal{M} \vDash \varphi[a]$ *&* $\mathcal{M} \vDash \psi[a]$,
   $\mathcal{M} \nvDash \varphi \wedge \psi[a] \Leftrightarrow \mathcal{M} \nvDash \varphi[a]$ *or* $\mathcal{M} \nvDash \psi[a]$;
6. $\mathcal{M} \vDash \varphi \otimes \psi[a] \Leftrightarrow \mathcal{M} \vDash \varphi[a]$ *&* $\mathcal{M} \vDash \psi[a]$,
   $\mathcal{M} \nvDash \varphi \otimes \psi[a] \Leftrightarrow \mathcal{M} \nvDash \varphi[a]$ *&* $\mathcal{M} \nvDash \psi[a]$;
7. $\mathcal{M} \vDash \forall x\, \varphi[a] \Leftrightarrow \mathcal{M} \vDash \varphi[a^x_d]$ *for all* $d \in \mathcal{D}$,
   $\mathcal{M} \nvDash \forall x\, \varphi[a] \Leftrightarrow \mathcal{M} \nvDash \varphi[a^x_d]$ *for some* $d \in \mathcal{D}$.

The following definition gives the connection between the elements of *FOUR* and combinations of truth and falsity.

Definition 8.2 *The value of a formula φ in a model $\mathcal{M}$ under an assignment a, $[\![\varphi]\!]^{\mathcal{M},a}$, is defined as follows.*

$$
\begin{array}{lll}
[\![\varphi]\!]^{\mathcal{M},a} = \mathbf{t} & \textit{iff} & \mathcal{M} \vDash \varphi[a] \textit{ and } \mathcal{M} \nvDash \varphi[a], \\
[\![\varphi]\!]^{\mathcal{M},a} = \mathbf{f} & \textit{iff} & \mathcal{M} \nvDash \varphi[a] \textit{ and } \mathcal{M} \nvDash \varphi[a], \\
[\![\varphi]\!]^{\mathcal{M},a} = \mathbf{n} & \textit{iff} & \mathcal{M} \nvDash \varphi[a] \textit{ and } \mathcal{M} \nvDash \varphi[a], \\
[\![\varphi]\!]^{\mathcal{M},a} = \mathbf{b} & \textit{iff} & \mathcal{M} \vDash \varphi[a] \textit{ and } \mathcal{M} \nvDash \varphi[a].
\end{array}
$$

In this paper we will restrict all discussion to sentences (closed formulas) and all mention of assignment functions will be dropped.

Definition 8.3 With Ξ and Θ sets of sentences of $\mathcal{L}$, we define the following relations.

- $\Xi \models^{tr} \Theta$ iff, for all models $\mathcal{M}$, $[\![\varphi]\!]^{\mathcal{M}} \in \{\mathbf{t}, \mathbf{b}\}$ for all $\varphi \in \Xi$ implies $[\![\psi]\!]^{\mathcal{M}} \in \{\mathbf{t}, \mathbf{b}\}$ for some $\psi \in \Theta$
- $\Xi \models^{nf} \Theta$ iff, for all models $\mathcal{M}$, $[\![\varphi]\!]^{\mathcal{M}} \in \{\mathbf{t}, \mathbf{n}\}$ for all $\varphi \in \Xi$ implies $[\![\psi]\!]^{\mathcal{M}} \in \{\mathbf{t}, \mathbf{n}\}$ for some $\psi \in \Theta$
- $\Xi \models^{f} \Theta$ iff, for all models $\mathcal{M}$, $[\![\varphi]\!]^{\mathcal{M}} \in \{\mathbf{f}, \mathbf{b}\}$ for all $\varphi \in \Xi$ implies $[\![\psi]\!]^{\mathcal{M}} \in \{\mathbf{f}, \mathbf{b}\}$ for some $\psi \in \Theta$
- $\Xi \models \Theta$ iff $\Xi \models^{tr} \Theta$ and $\Xi \models^{nf} \Theta$
- $\Xi \models\!\!\!\approx \Theta$ iff $\Xi \models^{tr} \Theta$ and $\Xi \models^{f} \Theta$

Entailment ($\models$) and necessary approximation ($\models\!\!\!\approx$) are the two relations of primary interest in this paper. Their relations to the orderings $\leq_t$ and $\leq_k$ were discussed in the introduction. They are derived relations, in the sense that $\models$ is defined as the preservation of truth and non-falsity, while $\models\!\!\!\approx$ is defined as preservation of truth and falsity. A syntactic characterisation of $\models$ can therefore be obtained by laying down proof rules corresponding to $\models^{tr}$ and $\models^{nf}$, while one for $\models\!\!\!\approx$ can be given by establishing proof rules corresponding to $\models^{tr}$ and $\models^{f}$. We will do so in our 'two trees' formalism, to be discussed in the next section.

As usual, the language will be extended with abbreviations and in fact all truth-functions are expressible since $\{\otimes, -, \wedge, \neg\}$ is functionally complete (see Muskens [37] for a proof). We will define $\vee$ and $\exists$ in the standard way. The following abbreviations may also be used.

Definition 8.4 *We will write*

$\mathbf{n}$	*for*	$-p \otimes \neg p$ *(where p is some fixed 0-place relation symbol);*
$\mathbf{b}$	*for*	$-\mathbf{n}$*;*
$\mathbf{f}$	*for*	$\mathbf{b} \wedge \mathbf{n}$*;*
$\mathbf{t}$	*for*	$\neg\mathbf{f}$*;*
$\varphi \oplus \psi$	*for*	$-(-\varphi \otimes -\psi)$*;*
$\varphi \mathbin{@} \psi$	*for*	$(\varphi \wedge \mathbf{b}) \vee (\psi \wedge \mathbf{n})$*;*
$\varphi \mathbin{/} \psi$	*for*	$(\varphi \wedge \psi) \mathbin{@} (\neg\varphi \vee \psi)$*;*
$\Pi x\varphi$	*for*	$\forall x\varphi \mathbin{@} \exists x\varphi$*;* *and*
$\Sigma x\varphi$	*for*	$\exists x\varphi \mathbin{@} \forall x\varphi$*.*

The first four zero-place connectives have the obvious denotation. The connective $\oplus$ is the natural dual of $\otimes$ and denotes join in A4. A sentence of the form $\varphi \mathbin{@} \psi$ is true iff φ is true and false iff ψ is false; $\varphi \mathbin{/} \psi$ is related to Blamey's [8] *transplication* and can be read as 'ψ, presupposing φ'. This formula has the value of ψ if φ is true, but is neither true nor false otherwise. The Π and Σ quantifiers are the duals of $\forall$ and $\exists$ and correspond to arbitrary meet and join in the approximation lattice A4. The operators /, Π and Σ will play no further role in this paper, but are interesting in their own right.

The proof system of the next section will be based on the four-sided sequents that were used in [38], following an idea described in Langholm [34]. Here is a pictorial representation of such a sequent.

$$\begin{array}{c|c} \Gamma_1 & \Delta_1 \\ \hline \Gamma_2 & \Delta_2 \end{array}\rightarrow$$

We linearise notation by attaching two *signs* i and j to formulas. i can be n *(north)* or s *(south)*, j can be e *(east)* or w *(west)*. The sequent displayed above will be written as

$$\{\varphi^{n,w} \mid \varphi \in \Gamma_1\} \cup \{\varphi^{n,e} \mid \varphi \in \Delta_1\} \cup \{\varphi^{s,w} \mid \varphi \in \Gamma_2\} \cup \{\varphi^{s,e} \mid \varphi \in \Delta_2\}.$$

The arrow in the picture is meant to signify transmission from left to right, meaning that whenever a model verifies all sentences in Γ_1 and falsifies all sentences in Γ_2 it must also either verify a sentence in Δ_1 or falsify a sentence in Δ_2. If this is not the case we say that the sequent is *refuted*$^{\rightharpoonup}$ by some model, a notion we define as follows.

Definition 8.5 *A model* $\mathcal{M}$ refutes$^{\rightharpoonup}$ *a sequent* Γ *if*

$$\begin{array}{rcl} \varphi^{n,w} \in \Gamma & \Longrightarrow & [\![\varphi]\!]^{\mathcal{M}} \in \{\mathbf{t}, \mathbf{b}\} \\ \varphi^{n,e} \in \Gamma & \Longrightarrow & [\![\varphi]\!]^{\mathcal{M}} \in \{\mathbf{f}, \mathbf{n}\} \\ \varphi^{s,w} \in \Gamma & \Longrightarrow & [\![\varphi]\!]^{\mathcal{M}} \in \{\mathbf{f}, \mathbf{b}\} \\ \varphi^{s,e} \in \Gamma & \Longrightarrow & [\![\varphi]\!]^{\mathcal{M}} \in \{\mathbf{t}, \mathbf{n}\} \end{array}$$

The dual notion is transmission from right to left, i.e. whenever a model falsifies all sentences in Δ_1 and verifies all sentences in Δ_2 it also falsifies a sentence in Γ_1 or verifies a sentence in Γ_2. The corresponding notion of *refutation*$^{\leftharpoondown}$ can be defined directly, but also in the following way.

Definition 8.6 *Let* Γ *be a sequent. We define the* dual *of* Γ, $\mathrm{dual}(\Gamma)$*, to be the sequent which results from* Γ *by simultaneously replacing every superscript* n *in* Γ *by* s*, every* s *by* n*, every* w *by* e*, and every* e *by* w*. A model* $\mathcal{M}$ refutes$^{\leftharpoondown}$ *a sequent* Γ *if* $\mathcal{M}$ refutes$^{\rightharpoonup}$ $\mathrm{dual}(\Gamma)$*.* $\mathcal{M}$ refutes Γ *if* $\mathcal{M}$ *either* refutes$^{\rightharpoonup}$ *or* refutes$^{\leftharpoondown}$ Γ*.*

$\Xi \models^{tr} \Theta$ iff there is no model refuting$^{\rightharpoonup}$ $\{\varphi^{n,w} \mid \varphi \in \Xi\} \cup \{\varphi^{n,e} \mid \varphi \in \Theta\}$, while $\Xi \models^{nf} \Theta$ iff no model refutes$^{\leftharpoondown}$ $\{\varphi^{n,w} \mid \varphi \in \Xi\} \cup \{\varphi^{n,e} \mid \varphi \in \Theta\}$, i.e. iff no model refutes$^{\rightharpoonup}$ $\{\varphi^{s,e} \mid \varphi \in \Xi\} \cup \{\varphi^{s,w} \mid \varphi \in \Theta\}$. Lastly, we have that $\Xi \models^{f} \Theta$ iff no model refutes$^{\rightharpoonup}$ $\{\varphi^{s,w} \mid \varphi \in \Xi\} \cup \{\varphi^{s,e} \mid \varphi \in \Theta\}$.

8.4 Proofs

Definition 8.7 *A sequent* has a proof tree *if it follows in the usual way from the following sequent rules. (We let* $-n = s, -s = n, -e = w, -w = e$*.)*

(R) $\quad \dfrac{}{\Gamma, \varphi^{i,w}, \varphi^{i,e}}$

(¬) $\quad \dfrac{\Gamma, \varphi^{i,j}}{\Gamma, \neg\varphi^{-i,j}}$

(−) $\quad \dfrac{\Gamma, \varphi^{i,j}}{\Gamma, -\varphi^{-i,-j}}$

($\wedge^{ne}_{sw}$*)* $\quad \dfrac{\Gamma, \varphi^{i,j} \quad \Gamma, \psi^{i,j}}{\Gamma, (\varphi \wedge \psi)^{i,j}}$, *where* $\langle i, j\rangle \in \{\langle n, e\rangle, \langle s, w\rangle\}$

$(\wedge^{nw}_{se})$ $\dfrac{\Gamma, \varphi^{i,j}, \psi^{i,j}}{\Gamma, (\varphi \wedge \psi)^{i,j}}$, *where* $\langle i,j\rangle \in \{\langle n,w\rangle, \langle s,e\rangle\}$

$(\otimes^{ne}_{se})$ $\dfrac{\Gamma, \varphi^{i,e} \quad \Gamma, \psi^{i,e}}{\Gamma, (\varphi \otimes \psi)^{i,e}}$

$(\otimes^{nw}_{sw})$ $\dfrac{\Gamma, \varphi^{i,w}, \psi^{i,w}}{\Gamma, (\varphi \otimes \psi)^{i,w}}$

$(\forall^{ne}_{sw})$ $\dfrac{\Gamma, [c/x]\varphi^{i,j}}{\Gamma, \forall x \varphi^{i,j}}$, *where* c *is not in* Γ *or* φ *and* $\langle i,j\rangle \in \{\langle n,e\rangle, \langle s,w\rangle\}$

$(\forall^{nw}_{se})$ $\dfrac{\Gamma, [t/x]\varphi^{i,j}}{\Gamma, \forall x \varphi^{i,j}}$, *where* $\langle i,j\rangle \in \{\langle n,w\rangle, \langle s,e\rangle\}$

(id) $\dfrac{}{\Gamma, t = t^{i,j}}$, *where* $\langle i,j\rangle \in \{\langle n,e\rangle, \langle s,w\rangle\}$

(L) $\dfrac{\Gamma, [t_2/x]\varphi^{i',j'}}{\Gamma, t_1 = t_2{}^{i,j}, [t_1/x]\varphi^{i',j'}}$, *where* $\langle i,j\rangle \in \{\langle n,w\rangle, \langle s,e\rangle\}$.

This calculus can be used to define several notions of entailment.

Definition 8.8 *We will write*

- $\Xi \vdash^{tr} \Theta$ *iff* $\{\varphi^{n,w} \mid \varphi \in \Xi\} \cup \{\varphi^{n,e} \mid \varphi \in \Theta\}$ *has a proof tree;*
- $\Xi \vdash^{nf} \Theta$ *iff* $\{\varphi^{s,e} \mid \varphi \in \Xi\} \cup \{\varphi^{s,w} \mid \varphi \in \Theta\}$ *has a proof tree;*
- $\Xi \vdash^{f} \Theta$ *iff* $\{\varphi^{s,w} \mid \varphi \in \Xi\} \cup \{\varphi^{s,e} \mid \varphi \in \Theta\}$ *has a proof tree;*
- $\Xi \vdash \Theta$ *iff* $\Xi \vdash^{tr} \Theta$ *and* $\Xi \vdash^{nf} \Theta$*;*
- $\Xi \mid\!\sim \Theta$ *iff* $\Xi \vdash^{tr} \Theta$ *and* $\Xi \vdash^{f} \Theta$.

In order to check that $\Xi \vdash \Theta$ it must clearly be shown that $\{\varphi^{n,w} \mid \varphi \in \Xi\} \cup \{\varphi^{n,e} \mid \varphi \in \Theta\}$ and its dual $\{\varphi^{s,e} \mid \varphi \in \Xi\} \cup \{\varphi^{s,w} \mid \varphi \in \Theta\}$ have proof trees. Here, for example, is one half of the proof that $\neg(\varphi \,/\, \psi) \vdash \varphi \,/\, \neg\psi$ (for all rules for defined symbols, see the Appendix):

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\varphi^{n,w},\ \psi^{s,w},\ \varphi^{n,e}}\,(R) \quad \dfrac{\dfrac{}{\varphi^{n,w},\ \psi^{s,w},\ \psi^{s,e}}\,(R)}{\varphi^{n,w},\ \psi^{s,w},\ \neg\psi^{n,e}}\,(\neg)}{\varphi^{n,w},\ \psi^{s,w},\ (\varphi \,/\, \neg\psi)^{n,e},}\,(/^{ne})}{(\varphi \,/\, \psi)^{s,w},\ (\varphi \,/\, \neg\psi)^{n,e}}\,(/^{sw})}{\neg(\varphi \,/\, \psi)^{n,w},\ (\varphi \,/\, \neg\psi)^{n,e}}\,(\neg)$$

And here is the other half:[7]

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{\varphi^{n,e},\ \varphi^{n,w},\ \psi^{n,w}}\,(R) \quad \dfrac{}{\psi^{n,e},\ \varphi^{n,w},\ \psi^{n,w}}\,(R)}{(\varphi\,/\,\psi)^{n,e},\ \varphi^{n,w},\ \psi^{n,w}}\,(/^{ne})}{(\varphi\,/\,\psi)^{n,e},\ \varphi^{n,w},\ \neg\psi^{s,w}}\,(\neg)}{(\varphi\,/\,\psi)^{n,e},\ (\varphi\,/\,\neg\psi)^{s,w}}\,(/^{sw})}{\neg(\varphi\,/\,\psi)^{s,e},\ (\varphi\,/\,\neg\psi)^{s,w}}\,(\neg)$$

This proof method, with each proof consisting of two proof trees instead of the usual single tree, is complete.

Theorem 8.1 *For all sets of sentences* Ξ *and* Θ*:*

1. $\Xi \vdash^{tr} \Theta \Longleftrightarrow \Xi \models^{tr} \Theta$
2. $\Xi \vdash^{nf} \Theta \Longleftrightarrow \Xi \models^{nf} \Theta$
3. $\Xi \vdash^{f} \Theta \Longleftrightarrow \Xi \models^{f} \Theta$
4. $\Xi \vdash \Theta \Longleftrightarrow \Xi \models \Theta$
5. $\Xi \mid\!\sim \Theta \Longleftrightarrow \Xi \mid\!\approx \Theta$

Proof. The proof rests upon the completeness proof given in [38]. That paper considers sequent calculi for a language containing the logical operators $\{\mathbf{n}, =, \neg, -, \wedge, \forall\}$. Here, $\mathbf{n}$ is a 0-place operator with the expected interpretation and $\{\mathbf{n}, \neg, -, \wedge, \}$, just as $\{\neg, \wedge, -, \otimes\}$, is functionally complete. In [38], a sequent is defined as a set of signed (as in this paper) formulae together with any subset of the set of *structural elements* $\{\not\vdash, \not\dashv\}$. The main sequent calculus that is considered ([38, Definition 6]) contains three sequent rules involving structural elements. However, [38, Remark 5.3] defines an alternative sequent calculus, called the *tr-calculus*, which is closely related to that of the present paper and which does away with structural elements. The tr-calculus consists of the following rules that are also present in the calculus of this paper:[8] (R), $(\neg)$, $(-)$, $\wedge^{ne}_{sw}$, $\wedge^{nw}_{se}$, $\forall^{ne}_{sw}$, $\forall^{nw}_{se}$, (id), and (L). Besides these familiar rules, the tr-calculus contains the following rule for $\mathbf{n}$:

$$(\mathbf{n}) \qquad \frac{}{\Gamma, \mathbf{n}^{i,w}}$$

First note that a sequent (in the sense of our paper) has a proof tree if and only if it is *tr-provable*, since $(\mathbf{n})$ is a derivable rule in our calculus and all rules in our calculus are at least admissible in the tr-calculus. It then follows by the results of [38] that a sequent Γ has a proof tree iff no model refutes$^{\rightharpoonup}$ it. The statements 1-5 follow easily from this. □

[7] In many cases it may not be necessary to expand a second proof tree. For example, if all formulas under consideration are classical, in the sense that they are built up using $\neg$, $\wedge$, $\forall$ and $=$ only, the second tree will be isomorphic to the first, as can easily be shown. Addition of $-$ will not change this; but addition of $\otimes$ or any of its ilk does.

[8] In [38], (R) and (L) were restricted to atomic formulae, while in the present paper this atomicity constraint is lifted.

Remark 1 As Remark 5.4 in [38] explains, the use of structural elements in that paper makes it possible to formulate rules for $\otimes$ and $\oplus$ without any violation of the subformula property. The present paper shows that a move to a 'two trees' system makes it possible to have the subformula property without structural elements.

8.5 Answer to a Question by Avron

In [3], Arnon Avron develops a unified tableau system, exploiting four signs, in terms of which sound and complete proof systems can be defined for various logics. One of the logics considered is an extension of Belnap's four valued logic with an appropriate implication connective, which is denoted as $\supset$ and defined as follows.

$$[\![\varphi \supset \psi]\!]^{\mathcal{M}_4} = \begin{cases} [\![\psi]\!]^{\mathcal{M}_4}, & \text{if } [\![\varphi]\!]^{\mathcal{M}_4} \in \{\mathbf{t}, \mathbf{b}\} \\ \mathbf{t}, & \text{if } [\![\varphi]\!]^{\mathcal{M}_4} \notin \{\mathbf{t}, \mathbf{b}\} \end{cases}$$

Here, $\mathcal{M}_4$ is a four valued model (sentences take values in $\{\mathbf{t}, \mathbf{f}, \mathbf{b}, \mathbf{n}\}$) for a propositional language, $\mathcal{L}_\star$, in the connectives $\{\neg, \wedge, \vee, \supset\}$. With the definition of $\supset$ just given, the definition of such a model can be left to the reader. The semantic consequence relation for $\mathcal{L}_\star$ that is considered by Avron is the preservation of truth (i.e., the values $\mathbf{t}$ and $\mathbf{b}$ are designated) as measured by $\mathcal{M}_4$ models; we denote this relation by $\models^{tr}_\star$. Avron's tableau system is shown to be sound and complete with respect to $\models^{tr}_\star$.

However, the connectives of the language $\mathcal{L}_\star$, $\{\neg, \wedge, \vee, \supset\}$, are, in contrast to the connectives $\{\neg, \wedge, -, \otimes\}$ that were considered in this paper, *not* truth functionally complete with respect to $\{\mathbf{t}, \mathbf{f}, \mathbf{b}, \mathbf{n}\}$. About the relation between his tableau system and connectives such as $-$ and $\otimes$, Avron asks the following question:

> For the Belnap logic, there is a second set of connectives that is sometimes considered (the knowledge / information ones). Can these be captured by tableau rules too? (Avron [3, p14])

Due to the close relation between sequent calculi and tableau systems, the results of this paper answer this question affirmatively. Table 1 gives tableau expansion rules for the connectives $\neg$, $-$, $\wedge$ and $\otimes$ that correspond closely to the Gentzen rules that were given before, but use Avron's [3] notation. Avron uses $\mathbf{T}^+$, $\mathbf{T}^-$, $\mathbf{F}^+$ and $\mathbf{F}^-$ in order to sign sentences; here + corresponds to our n, $-$ to s, $\mathbf{T}$ to w, and $\mathbf{F}$ to e.

Together with the obvious closure condition (a branch is closed if it contains either $\{\mathbf{T}^+\varphi, \mathbf{F}^+\varphi\}$ or $\{\mathbf{T}^-\varphi, \mathbf{F}^-\varphi\}$) this readily gives characterisations of $\models^{tr}$, $\models^{nf}$, $\models^{f}$, $\models$, and $\models\!\!\!\!\sim$ on the propositional fragment of the language (rules for the quantifiers can easily be added). In order to check whether $\Xi \models^{tr} \Theta$, for example, a tableau for $\{\mathbf{T}^+\varphi \mid \varphi \in \Xi\} \cup \{\mathbf{F}^+\varphi \mid \varphi \in \Theta\}$ should be expanded, while checking whether $\Xi \models^{nf} \Theta$ requires expansion of $\{\mathbf{F}^-\varphi \mid \varphi \in \Xi\} \cup \{\mathbf{T}^-\varphi \mid \varphi \in \Theta\}$ and checking whether $\Xi \models \Theta$ in general requires both.

The tableau system given here properly extends Avron's. It extends it because the rules for $\neg$ and $\wedge$ given here correspond to Avron's rules, Avron's rules for $\vee$ are derivable, and his rules for $\supset$ are derivable once $\varphi \supset \psi$ is taken to be

$$\frac{\mathbf{T}^+\neg\varphi}{\mathbf{T}^-\varphi} \qquad \frac{\mathbf{T}^-\neg\varphi}{\mathbf{T}^+\varphi} \qquad \frac{\mathbf{F}^+\neg\varphi}{\mathbf{F}^-\varphi} \qquad \frac{\mathbf{F}^-\neg\varphi}{\mathbf{F}^+\varphi}$$

$$\frac{\mathbf{T}^+ - \varphi}{\mathbf{F}^-\varphi} \qquad \frac{\mathbf{T}^- - \varphi}{\mathbf{F}^+\varphi} \qquad \frac{\mathbf{F}^+ - \varphi}{\mathbf{T}^-\varphi} \qquad \frac{\mathbf{F}^- - \varphi}{\mathbf{T}^+\varphi}$$

$$\frac{\mathbf{T}^+\varphi \wedge \psi}{\mathbf{T}^+\varphi,\ \mathbf{T}^+\psi} \qquad \frac{\mathbf{T}^-\varphi \wedge \psi}{\mathbf{T}^-\varphi \mid \mathbf{T}^-\psi} \qquad \frac{\mathbf{F}^+\varphi \wedge \psi}{\mathbf{F}^+\varphi \mid \mathbf{F}^+\psi} \qquad \frac{\mathbf{F}^-\varphi \wedge \psi}{\mathbf{F}^-\varphi,\ \mathbf{F}^-\psi}$$

$$\frac{\mathbf{T}^+\varphi \otimes \psi}{\mathbf{T}^+\varphi,\ \mathbf{T}^+\psi} \qquad \frac{\mathbf{T}^-\varphi \otimes \psi}{\mathbf{T}^-\varphi,\ \mathbf{T}^-\psi} \qquad \frac{\mathbf{F}^+\varphi \otimes \psi}{\mathbf{F}^+\varphi \mid \mathbf{F}^+\psi} \qquad \frac{\mathbf{F}^-\varphi \otimes \psi}{\mathbf{F}^-\varphi \mid \mathbf{F}^-\psi}$$

Table 8.1: Expansion rules for propositional connectives.

an abbreviation of $\neg(\varphi \mathbin{@} -\varphi) \vee \psi$.[9] The extension is proper, as $\{\neg, \wedge, -, \otimes\}$ are functional complete while [2, Theorem 14] shows that $\{\neg, \wedge, \vee, \supset\}$ is not.

While our tableau system characterising $\models^{tr}$ thus extends Avron's system characterising $\models^{tr}_{\star}$, we feel that the entailment relation that correctly captures the spirit of Belnap's logic, the one in which entailment corresponds with $\leq_t$, is $\models$, not $\models^{tr}$ (see also footnote 6).

8.6 Conclusion

We have shown how Belnap's logic can be provided with an analytic Gentzen calculus that is completely natural. The price is that, in general, every proof now comes with *two* proof trees instead of one. While this idea may seem strange at first, it fits well with the observation that doubling of concepts is a general phenomenon in Belnap's logic.

8.7 Appendix: Gentzen Rules for Defined Operators

$(\mathbf{n})$ $\quad \dfrac{}{\Gamma, \mathbf{n}^{i,w}}$

$(\vee^{nw}_{se})$ $\quad \dfrac{\Gamma, \varphi^{i,j} \quad \Gamma, \psi^{i,j}}{\Gamma, (\varphi \vee \psi)^{i,j}}$, where $\langle i,j\rangle \in \{\langle n,w\rangle, \langle s,e\rangle\}$

$(\vee^{ne}_{sw})$ $\quad \dfrac{\Gamma, \varphi^{i,j}, \psi^{i,j}}{\Gamma, (\varphi \vee \psi)^{i,j}}$, where $\langle i,j\rangle \in \{\langle n,e\rangle, \langle s,w\rangle\}$

$(\exists^{nw}_{se})$ $\quad \dfrac{\Gamma, [c/x]\varphi^{i,j}}{\Gamma, \exists x\varphi^{i,j}}$, where c is not in Γ or φ and $\langle i,j\rangle \in \{\langle n,w\rangle, \langle s,e\rangle\}$

$(\exists^{ne}_{sw})$ $\quad \dfrac{\Gamma, [t/x]\varphi^{i,j}}{\Gamma, \exists x\varphi^{i,j}}$, where $\langle i,j\rangle \in \{\langle n,e\rangle, \langle s,w\rangle\}$

[9] Note that this formula has the right semantics, as $\varphi \mathbin{@} -\varphi$ gets the value $\mathbf{t}$ if φ has a value in $\{\mathbf{t}, \mathbf{b}\}$ and gets the value $\mathbf{f}$ otherwise.

$(@^{ne}_{nw})$ $\dfrac{\Gamma, \varphi^{n,j}}{\Gamma, (\varphi @ \psi)^{n,j}}$

$(@^{se}_{sw})$ $\dfrac{\Gamma, \psi^{s,j}}{\Gamma, (\varphi @ \psi)^{s,j}}$

$(\oplus^{ne}_{se})$ $\dfrac{\Gamma, \varphi^{i,e}, \psi^{i,e}}{\Gamma, (\varphi \oplus \psi)^{i,e}}$

$(\oplus^{nw}_{sw})$ $\dfrac{\Gamma, \varphi^{i,w} \quad \Gamma, \psi^{i,w}}{\Gamma, (\varphi \oplus \psi)^{i,w}}$

$(/^{ne})$ $\dfrac{\Gamma, \varphi^{n,e} \quad \Gamma, \psi^{n,e}}{\Gamma, (\varphi \,/\, \psi)^{n,e}}$

$(/^{se})$ $\dfrac{\Gamma, \varphi^{n,e} \quad \Gamma, \psi^{s,e}}{\Gamma, (\varphi \,/\, \psi)^{s,e}}$

$(/^{nw})$ $\dfrac{\Gamma, \varphi^{n,w}, \psi^{n,w}}{\Gamma, (\varphi \,/\, \psi)^{n,w}}$

$(/^{sw})$ $\dfrac{\Gamma, \varphi^{n,w}, \psi^{s,w}}{\Gamma, (\varphi \,/\, \psi)^{s,w}}$

(Π^{ne}_{se}) $\dfrac{\Gamma, [c/x]\varphi^{i,e}}{\Gamma, \Pi x \varphi^{i,e}}$
(c not in Γ or φ)

(Π^{nw}_{sw}) $\dfrac{\Gamma, [t/x]\varphi^{i,w}}{\Gamma, \Pi x \varphi^{i,w}}$

(Σ^{ne}_{se}) $\dfrac{\Gamma, [t/x]\varphi^{i,e}}{\Gamma, \Sigma x \varphi^{i,e}}$

(Σ^{nw}_{sw}) $\dfrac{\Gamma, [c/x]\varphi^{i,w}}{\Gamma, \Sigma x \varphi^{i,w}}$
(c not in Γ or φ)

Bibliography

[1] O. Arieli and A Arnon. Reasoning with Logical Bilattices. *Journal of Logic Language and Information*, 5:25–63, 1996.

[2] O. Arieli and A. Avron. The Value of the Four Values. *Artificial Intelligence*, 102:97–141, 1998.

[3] A. Avron. Tableaux with Four Signs as a Unified Framework. In *TABLEAUX*, pages 4–16, 2003.

[4] M. Baaz, C.G. Fermüller, and G. Salzer. Automated Deduction for Many-Valued Logics. In A. Robinson and A. Voronkov, editors, *Handbook of Automated Reasoning*, pages 1355–1402. Elsevier Science Publishers, 2000.

[5] J. Beall. *Spandrels of Truth*. Oxford University Press, 2009.

[6] N.D. Belnap. How a Computer Should Think. In G. Ryle, editor, *Contemporary Aspects of Philosophy*, pages 30–56. Oriel Press, Stocksfield, 1976.

[7] N.D. Belnap. A Useful Four-Valued Logic. In J.M. Dunn and G.Epstein, editors, *Modern Uses of Multiple-Valued Logic*. 1977.

[8] S. Blamey. Partial Logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume III, pages 1–70. Reidel, Dordrecht, 1986.

[9] G. Boolos. The Hardest Logic Puzzle ever. *The Harvard Review of Philosophy*, 6:62–65, 1996.

[10] F. Buekens. Truth's Role in Understanding, in the Analytical Way. 2009.

[11] P. Cobreros, P. Egré, D. Ripley, and R. van Rooij. Tolerant, Classical, Strict. *Journal of Philosophical Logic (to appear)*, 2011.

[12] R. de Wolf. *Quantum Computing and Communication Complexity*. PhD thesis. ILLC dissertation series, 2001.

[13] J.M. Dunn. Intuitive Semantics for First-Degree Entailments and 'Coupled Trees'. *Philosophical Studies*, 29:149–168, 1976.

[14] H. Field. Tarski's Theory of Truth. *Journal of Philosophy*, 69:347–375, 1972.

[15] H. Field. *Saving Truth from Paradox*. Oxford University Press, 2008.

[16] M. Fitting. Notes on the Mathematical Aspects of Kripke's Theory of Truth. *Notre Dame Journal of Formal Logic*, 27:75–88, 1986.

[17] M. Fitting. Bilattices and the Semantics of Logic Programming. *Journal of Logic Programming*, 11:91–116, 1991.

[18] H. Gaifman. Pointers to Truth. *Journal of Philosophy*, 89 (5):223–261, 1992.

[19] P. Geach. Assertion. *The Philosophical Review*, 74:449–465, 1965.

[20] M.L. Ginzberg. Multivalued Logics: A Uniform Approach to Reasoning in AI. *Computer Intelligence*, 4:256–316, 1988.

[21] D. Grover. How Significant is the Liar? In J. Beall and B. Armour-Garb, editors, *Deflationsim and Paradox*. 2005.

[22] L.K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings, 28th Annual ACM Symposium on the Theory of Computing*. 1996.

[23] A. Gupta. Truth and Paradox. *Journal of Philosophical Logic*, 11:1–60, 1982.

[24] A. Gupta and N. Belnap. *The Revision Theory of Truth*. MIT Press, Cambridge, 1993.

[25] J. Henson. The Labyrinth. 1986.

[26] J. Hintikka. *Revisiting the Principles of Mathematics*. Cambridge University Press, New York, 1996.

[27] L. Horsten. Levity. *Mind*, 118:555–581, 2009.

[28] P. Horwich. *Truth, second edition*. Oxford University Press, 1999.

[29] M. Kremer. Kripke and the Logic of Truth. *Journal of Philosophical Logic*, 17:225–278, 1988.

[30] P. Kremer. On the 'Semantics' for Languages with their own Truth Predicates. In A. Chapuis and A. Gupta, editors, *Truth, Defintion and Circularity*, pages 217–246. Indian Council of Philosophical Research, New Dehli, 2000.

[31] P. Kremer. Comparing Fixed-Point and Revision Theories of Truth. *Journal of Philosophical Logic*, 38:363–403, 2009.

[32] P. Kremer. How Truth Behaves When There's No Vicious Reference. *Journal of Philosophical Logic*, 39:345–367, 2010.

[33] S. Kripke. Outline of a Theory of Truth. *Journal of Philosophy*, 72:690–716, 1975.

[34] T. Langholm. How Different is Partial Logic? In P. Doherty, editor, *Partiality, Modality, and Nonmonotonicity*, pages 3–43. CSLI, Stanford, 1996.

[35] H. Leitgeb. What is a self-referential sentence? critical remarks on the alleged (non-) circularity of Yablo's paradox. *Logique & Analyse*, 177-178:3–14, 2002.

[36] P. Lorenzen and K. Lorenz. *Dialogische Logik*. Wissenschaftliche Buchgesellschaft, 1978.

[37] R.A. Muskens. *Meaning and Partiality*. CSLI, Stanford, 1995.

[38] R.A. Muskens. On Partial and Paraconsistent Logics. *Notre Dame Journal of Formal Logic*, 40:352–373, 1999.

[39] M.A. Nielsen and I.L. Chang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[40] G. Priest. Yablo's paradox. *Analysis*, 57:236–242, 1997.

[41] G. Priest. *In Contradiction (second edition)*. Oxford University Press, 2006.

[42] H. Putnam. *Meaning and the Moral Sciences*. Routledge and Kegan Paul, 1978.

[43] B. Rabern and L. Rabern. A Simple Solution to the Hardest Logic Puzzle Ever. *Analysis*, 68:105–112, 2008.

[44] B. Rabern and L. Rabern. In Defense of the Two Question Solution to the Hardest Logic Puzzle Ever. *Unpublished*, 2009.

[45] D. Ripley. Conservatively Extending Classical Logic with Transparant Truth. *Submitted*, 201x.

[46] D. Ripley. Paradox and Failures of Cut. *Australasian Journal of Philosophy (To appear)*, pages 1–25, 201x.

[47] T. Roberts. Some Thoughts about the Hardest Logic Puzzle Ever. *Journal of Philosophical Logic*, 30:609–612, 2001.

[48] I. Rumfitt. 'Yes' and 'No'. *Mind*, 109 (436):781–823, 2000.

[49] R. Smullyan. *The Lady or the Tiger*. Pelican Books, 1983.

[50] R Smullyan. *First-order Logic*. Dover, New York, 1995.

[51] Z. G. Szabó. Compositionality. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Winter 2008 Edition)*. 2008.

[52] R. Urbaniak. Leitgeb, "about" Yablo. *Logique & Analyse*, 207:239–254, 2009.

[53] G. Uzquiano. How to Solve the Hardest Logic Puzzle Ever in Two Questions. *Analysis*, 70:39–44, 2010.

[54] B. van Fraassen. *Laws and Symmetry*. Oxford University Press, 1989.

[55] M. Weiner. Must We Know What We Say? *Philosophical Review*, 114:227–251, 2005.

[56] G. Wheeler and P. Barahona. Why the Hardest Logic Puzzle Ever cannot be Solved in Less than Three Questions. *Journal of Philosophical Logic*, 2011.

[57] S. Wintein. Assertoric Semantics and the Computational Power of Self-Referential Truth. *Journal of Philosophical Logic (Section 4 of this thesis)*, 2011.

[58] S. Wintein. A Framework for Riddles about Truth that do not involve Self-Reference. *Studia Logica (Section 2 of this thesis)*, 98 (3):445–482, 2011.

[59] S. Wintein. What makes a knight? *Interfaces: Explorations in Logic, Language and Computation Lecture Notes in Computer Science*, 6211/2010:25–37, 2011.

[60] S. Wintein. Alternative Ways for Truth to Behave when there is no Vicious Reference. *Resubmitted to the Journal of Philosophical Logic. (Section 6 of this thesis)*, 201x.

[61] S. Wintein. Circularity and Paradoxality. *In Preparation*, 201x.

[62] S. Wintein. English and Prussian Assertibility in Languages of Self-Referential Truth. *In Preparation*, 201x.

[63] S. Wintein. From Closure Games to Generalized Strong Kleene Theories of Truth. *Submitted (Section 5 of this thesis)*, 201x.

[64] S. Wintein. On Languages that Contain their own Ungroundedness Predicate. *To appear in: Logique et Analyse*, 201x.

[65] S. Wintein. On the Behavior of True and False. *To appear in: Minds and Machines (Section 3 of this thesis)*, 201x.

[66] S. Wintein. Strict-Tolerant Tableaux for Strong Kleene Truth. *Submitted (Section 7 of this thesis)*, 201x.

[67] S. Wintein and R.A. Muskens. A Calculus for Belnap's Logic in Which Each Proof Consists of Two Trees. *Logique & Analyse (Section 8 of this thesis)*, 201x.

[68] L. Wittgenstein. *Lectures on the Foundations of Mathematics, Cambridge 1939*. Hassocks, Harvestor, ed. C. Diamond, 1939.

[69] S. Yablo. Paradox without Self-Reference. *Analysis*, 53:251–252, 1993.